Small regulatory RNAs

Seinen, Erwin

Small Regulatory RNAs: Identification, Classification and Utilization

Erwin Seinen

Colofon

Small Regulatory RNAs: Identification, Classification and Utilization.
Proefschrift, Rijksuniversiteit Groningen, Nederland

The research in this thesis was performed at the Department of Cell Biology, Section Radiation and Stress Cell Biology, University Medical Center Groningen, The Netherlands.

Cover: Erwin Seinen
Lay-out: Erwin Seinen
Helix picture: Jordan Edgcomb

The printing of this thesis was financially supported by:

- University of Groningen
- E. Seinen Beheer BV
- Acera Accountants

RIJKSUNIVERSITEIT GRONINGEN

Small Regulatory RNAs: Identification, Classification and Utilization

Proefschrift

ter verkrijging van het doctoraat in de
Medische Wetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op
maandag 11 april 2011
om 16.15 uur

door

Erwin Seinen

geboren op 28 februari 1978
te Leeuwarden

Promotores:	Prof. dr. O.C.M. Sibon Prof. dr. R.C. Jansen
Beoordelingscommissie:	Prof. G. de Haan (UMCG) Prof. J.J. Schuringa (UMCG) Prof. J.H.M. van den Berg (UMCG)

ISBN 978-90-367-4838-4 (gedrukte versie)

ISBN 978-90-367-4837-7 (digitale versie)

Contents

Chapter 1	Introduction	7
Chapter 2	A Genome-Wide analysis of the specificity of RNAi in Drosophila melanogaster using the novel tool RNAiSelect	23
Chapter 3	RNAi experiments in D. melanogaster: solutions to the overlooked problem of off-targets shared by independent dsRNAs	53
Chapter 4	Pantethine rescues a Drosophila model for pantothenate kinase–associated neurodegeneration	83
Chapter 5	A high throughput experimental approach to identify miRNA targets in human cells	117
Chapter 6	Summarizing discussion	151
Chapter 7	Nederlandse Samenvatting	161
Dankwoord		172
Curriculum Vitae		175

Chapter 1

Introduction

Introduction

Determining the function of genes and the complex regulatory interactions between them is an important objective within Life Science research. Before, the function of specific genes was primarily studied by inducing genetic mutations in targeted genes of interest. After the disruption of a specific gene, the to-be-translated functional protein will never be produced. These so called 'knock-out' studies allow the researcher to deduce the original function of the subjected gene, because the observed phenotypes are a consequence of the absence of a specific protein. A major drawback in classic knock-out studies using gene mutations is that it is relatively difficult to target specifically the genes of interest and it requires at least 1 or more generations before the model-organism has been designed and is ready for experiments. In addition, this type of experiments are rather crude because the gene is knocked out throughout the life cycle of the model-organisms which may have collateral effects not directly related to the knocked-out gene of interest. The RNAi (RNA Interference) technology has enriched and accelerated knock-out studies because it allows interfering at the RNA level [1]. This technique allows quantitative modification of RNA levels at any time and place of most organisms in a wild type background without the necessity to go through multiple generations of progeny of the model-organisms. By downregulating the mRNAs of the gene of interest, the functioning of the gene itself is abolished. This type of study allows straightforward gene expression manipulation and is also referred to as 'knock-down' study. RNAi-based research has led to the discovery of the microRNA (miRNA) pathway, which is involved in endogenous mechanisms of gene regulation. Multiple examples exist in which altered expression/activities of miRNAs are associated with pathologies.

The RNAi and miRNA pathway

RNAi exists as double stranded RNA of variable length which can enter an endogenous pathway known as the RNAi pathway. The RNAi pathway is responsible for the interference of messenger RNA (mRNA) translation based on sequence homology between the mRNA and the RNAi molecule. This thesis focuses on long double stranded RNA (dsRNA) of about 350 nucleotides (nt) in length, short interfering RNA (siRNA) of about 20-25 nt in length, and microRNAs (miRNAs) of about 20-25 nt in length.

In *Drosophila melanogaster*, RNAi is achieved by introducing an artificial dsRNA containing a sequence that is homologous to the gene of interest into cells, tissues or whole organisms to silence the translation of mRNAs of the targeted gene. By the action of the endogenous nuclease Dicer, this dsRNA is first cleaved into shorter fragments of about 20-25 nucleotides [2] and are referred to as the siRNAs. The siRNAs then split into single strand RNAs, of which one (the so called guide strand) is integrated into a protein complex known as the RNA-induced silencing complex (RISC) [3]. Although there are indications for a predominant determinant of which RNA strand will become the guide strand, other reports show that both are equally eligible for assembly into Argonaute, the catalytic component of the RISC complex [4, 5]. The RISC complex finds mRNAs that contain complementary sequences to the guide strand and actively cleaves the mRNA by the action of Argonaute which prevents translation. The siRNAs may also be included in the RNA-induced transcriptional silencing (RITS) complex. This complex can trigger posttranslational modification of histones resulting in a more permanent state of transcriptional silencing of the target gene [6]. This type of silencing is also referred to as epigenetics (see Figure 1 for an illustration of the different mechanisms) [7].

MicroRNAs (miRNAs) are produced from endogenous RNA encoding regions that regulate gene expression in a similar way as exogenous introduced RNAi molecules. The primary transcripts are processed within the nucleus to form a stem-loop structure called the pre-miRNA. The pre-miRNAs are cleaved by Dicer and therefore miRNA-induced events share much of the same downstream machinery as is controlled by Dicer and the RISC complex, but pre-miRNAs require some pre-modifications which is done by the microprocessor complex [8]. The miRNAs originate from noncoding regions and from intronic sequences [9]. They require limited homology to the target genes for being effective; only a stretch of 6-7 nt with 100% homology is sufficient and is denoted as the ‘seed’ region of the miRNA[10].

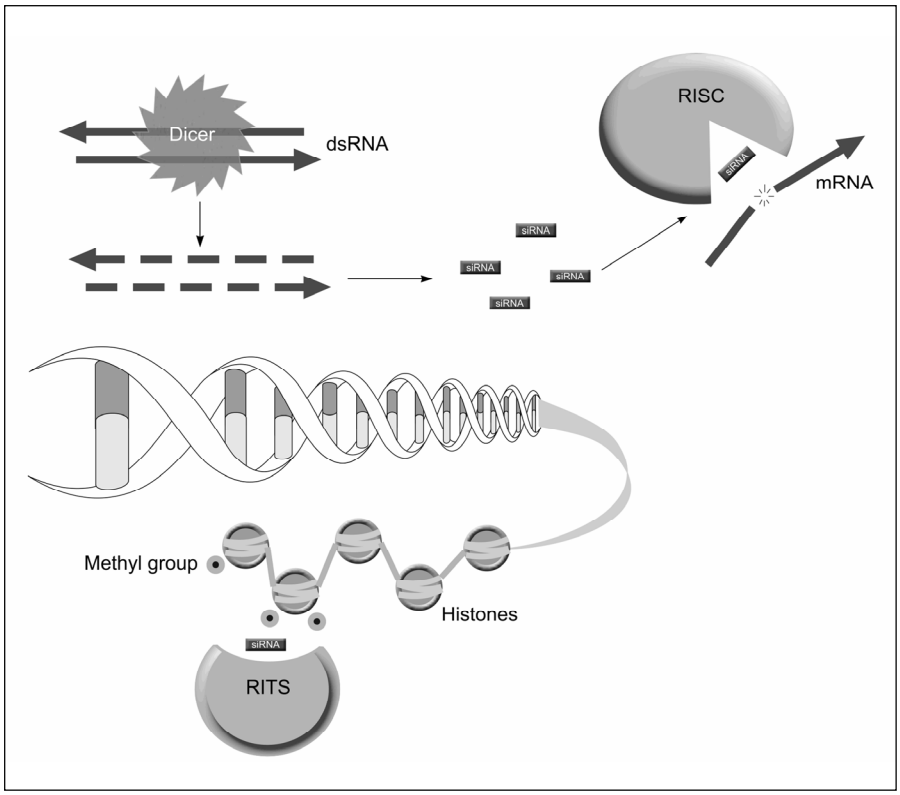


Figure 1 – Illustration of RNAi mechanisms. Dicer cuts the dsRNA into smaller siRNAs. The siRNAs is then either incorporated into RISC for RNA degradation or RITS for histone methylation.

Together, RNAs have been discovered as being very important in gene expression regulation, in a much more fine-tuned manner than the well-known gene promoter on-off actions [11]. RNAs are part of the regulatory system of living organisms (endogenously through miRNAs) and are effectively being exploited to influence the expression of the mRNA from targeted genes (exogenously through RNAi technology).

The specificity of knock-down studies

The biological results of the experiment in which the expression of specific genes are knocked down through RNAi can be difficult to interpret, because collateral effects can be present obscuring the relevant phenotype caused by the gene under study. These collateral effects are known as off-targets and are caused by (1) cellular responses due to the penetration of exogenous

RNA, and (2) through sequence homology of an RNAi construct with other genes than its intended target. Fortunately much has been discovered about the cellular responses, for example the interferon response which is triggered by double stranded RNA viruses [12]. By using strict design policies, and using proper transformation agents, these types of effects from (1) may be minimised or prevented altogether. However, off-targets caused by sequence homology (2) are still a challenge today [13-21]which will be discussed in more detail next.

An RNAi molecule is not exclusively specific because mismatches can be tolerated [10, 22] and not the complete length of an RNAi molecule is required to bind to mRNAs to induce a measurable effect [17, 18]. Birmingham et al. and Jackson et al. did a detailed analysis in human cell-based models and found that near-perfect matches (for example 18 matches throughout the whole siRNA) are sufficient to induce a change in protein expression [13, 17]. They, as well as others, also found that imperfect matches, apart from the 'seed' region of 6-nt at the 3' UTR (see further below) may contribute to off-targeting. Ma et al., Kulkarni et al. and Moffat et al. have shown that in *Drosophila*, many dsRNAs show off-target effects due to imperfect matches, as well as so called CAN repeats (N indicates any base) in low complex sequences within the dsRNA [19, 21]. Together their results reveal the strong evidence for false-positive rates in RNA interference (RNAi) screens using long dsRNAs.

It is also known that RNAi may target regions other than mature mRNAs, which greatly expands the possible unwanted targets to which RNAi may act upon [23-27]. The most relevant facts were found by Matzke et al., Bosher et al and Robb et al. who showed that RNAi can act within the nucleus, therefore other RNA molecules than mature mRNA can be influenced such as pre-mRNA containing intronic sequences.

Although still a challenge, bioinformatics may be of great help in analysing the genome and determining the specificity of RNAi molecules. As this thesis will show, this approach can help preventing many of the off-target effects which are caused by sequence similarities to non-targeted sequences.

Challenges for bioinformatics tools

Computers are a necessity during the design of RNAi constructs due to the large genomic data which need to be processed. Tools are available that

facilitate the search for siRNA candidates which have certain characteristics proven to show the best knock-down efficiency for the gene of interest (see table 1 for a comparison). This is done by gathering all known characteristics of identified highly potential siRNAs such as CG content and thermodynamic properties, and extrapolating this information onto the genome to find other highly active siRNAs [28]. Other tools are available that analyse siRNAs to find sequence similarities throughout the genome that might induce off-target effects [29].Using the output of this off-target analysis, the most specific siRNA can be chosen amongst the many that are predicted to be potentially active.

Together, there are tools available that all have their own strength and weaknesses (see table 1 for a comparison). Unfortunately this has also led to misunderstandings. For example, BLAST (Basic Local Alignment Tool) is a very popular tool to perform all kinds of alignments. It is readily available through an on-line interface and is extremely fast in generating alignment reports. Because finding off-targets for a particular siRNA involves reviewing possible alignments against the genome, BLAST was also quickly adopted for this kind of work. However, the strength of BLAST lies particularly in large-to-large sequence alignments and is less useful for alignment studies of short sequences such as most siRNAs. BLAST uses a so called 'word size' that dictates the minimum contiguous sequence homology before any positive hit may be found. The on-line version of BLAST has a minimum word-size parameter of 7, which is necessary for the algorithm to perform within a timely fashion and allowing on-line data management. This minimum word-size of BLAST becomes a problem if a 21-nt sequence has for example 3 mismatches when aligned to a potential off-target sequence on the genome. These numbers of mismatches within a sequence can still provoke a potent RNAi response towards a complementary mRNA [19, 20, 22]. This complementary mRNA can be considered as an off-target effect. In case 3 mismatches are evenly spread across a 21-nt sequence, BLAST would require a minimum word size of 21/(3+1)=5.25 to find this particular off-target. In other words, BLAST will miss this potential off-target with a word size of 7. This possibility is illustrated in Figure 2.

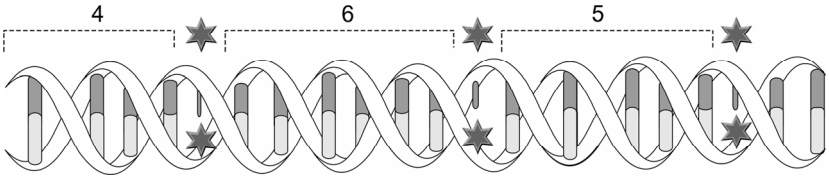


Figure 2 - *Schematic representation of 2 single stranded RNAs that complementary bind with the exception of 3 positions (mismatches). The stars represent a mismatch due to incompatible nucleotides. This figure demonstrates in this example that the largest contiguous stretch of homology to be found is 6. The online version of BLAST however, requires at least 7 contiguous matches for a positive match. Blast will not identify this off-target, indicating that BLAST is not suitable for accurate off-target analysis.*

There are tools that have smaller word sizes or have other methods for a more accurate alignment [28, 29]. For example there are tools available that are based on the Smith-Waterman algorithm which offers highly detailed alignment results [30]. However due to the long processing time, these types of algorithm are mostly used off-line with specialized computer systems [31-33]. Although it is possible to perform genome-wide analyses using these off-line systems, they take longer to complete than what is suitable for large scale and high-throughput siRNA analysis. Moreover the complexity of setting up these systems and lack of on-line availability makes them unreachable for non-bioinformaticians who do have the need for genome-wide off-target analysis.

Other tools are available that have different, more efficient algorithms that offer high-speed analysis without compromising sensitivity [28, 29]. These tools gain much of their acceleration by reducing the genome to non-redundant sequences that only include mature RNA (mRNA) sequences. This reduces the amount of nucleotides, which needs to be scanned, to a great extent. This in turn results in a significant increase of the performance and in creating the possibilities for widely adoptable online tools or high-throughput analysis. The general idea behind this approach was that RNAi is believed to be exclusively active within the cytoplasm and therefore will act upon mRNA sequences only [34]. Based on this assumption, it would indeed be sufficient to search only within mRNA sequences. However, research papers have reported evidence about RNAi being active within the nucleus, indicating that the previous assumption is not correct [23-27].

Chapter 2 of this thesis presents a novel tool called RNAi-Select to reduce off-target effects by performing highly sensitive sequence comparisons using a novel algorithm that is more comprehensive and appropriate than existing tools. By generating a comprehensive sequence alignment report, the user is able to choose a particular RNAi molecule with high specificity and by using this selected RNAi molecule off-targets can be prevented as much as possible.

	RNAi-Select	**BLAST**	**dsCheck / siDirect**
Freely on-line available application	Yes	Yes	Yes
Speed	Fast	Fast	Fast
Accurate short sequence alignment	Yes	**No**	Yes
Finds 1 (G:U or other) mismatches	Yes	Yes	Yes
Finds 2 (G:U or other) mismatches	Yes	**No**	Yes
Finds 3+ (G:U or other) mismatches	Yes	**No**	**No**
Finds exon or UTR based siRNA off-targets	Yes	**Some**	Yes
Finds whole genome siRNA off-targets, including introns	Yes	**Some**	**No**
Finds seed based miRNA off-targets	Yes	**No**	**No**
Designs specific shRNAs to knockdown *Drosophila* genes[1]	Yes	**No**	**No**
Average validated off-targets found per dsRNA[2]	13	**0**	2
Identification of non-overlapping dsRNAs with no shared off-targets	Yes	**No**	**No**

Table 1 – *Comparison and features of RNAi-Select and two other on-line available tools*

A double controlled strategy to circumvent off-targets

In many published manuscripts, two independent RNAi constructs have been used in separate experiments to downregulate the gene of interest [35-37], because previously it was demonstrated that this strategy allows to filter for off-target effects [20, 38]. The assumption is that when two non-overlapping RNAi constructs induce the same phenotype, this phenotype is an on-target effect and most likely not an off-target effect. This is reasonable because any phenotypical observation that occur in both dsRNA experiments is expected to be the result of downregulation of the intended target and not from any off-targets. Because it is also reasonable to assume

that each dsRNA has its own unique set of possible off-targets. In other words, any induced effect that occurs only after the use of one dsRNA both not in the other is attributed to off-target effects of that particular dsRNA and should be disregarded. Again, this postulation is based on the assumption that each dsRNA has a unique off-target profile, however this has never been tested experimentally. One major flaw in this assumption is that an off-targeted mRNA does not exist of a short 21-nt fragment, but is in itself most of the time a sequence of more than 400 nt with a large active surface having many possibilities of having sequence similarities to siRNAs. Thus even though different non-overlapping dsRNAs may have a completely different sequence composition, in theory their derived siRNAs can still bind to the same off-target mRNA albeit at different sites. Figure 3 presents an illustration of this possible event.

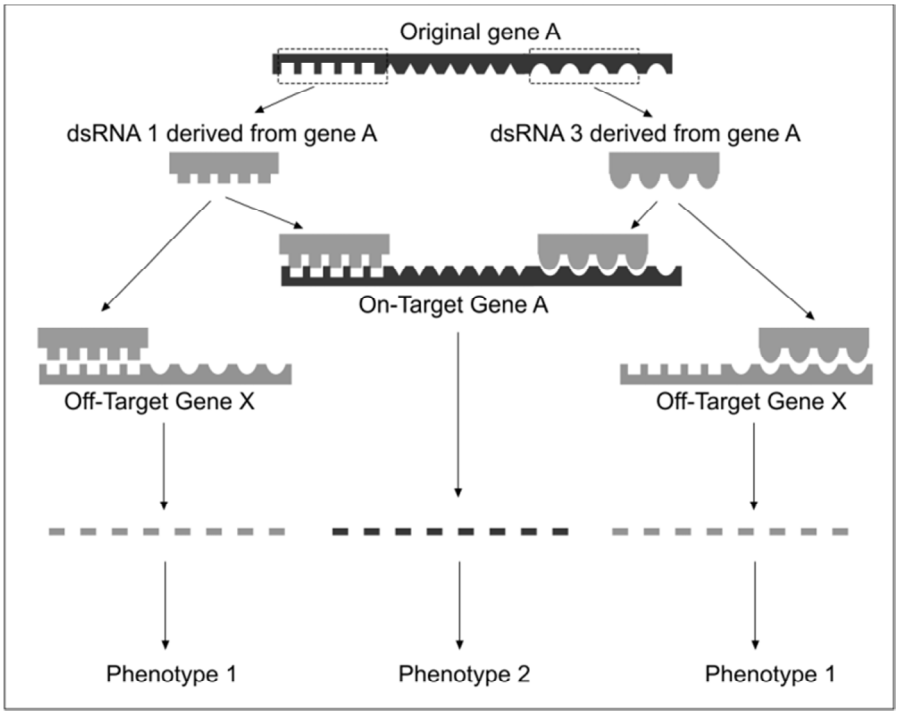


Figure 3 – Schematic representation of 2 non-overlapping dsRNAs that have sequence similarity to the same off-target.

Chapter 3 presents evidence that non-overlapping sequences derived from the same gene have an almost 100% change of sharing identical potential off-targets. We present a proper approach to select multiple dsRNA

molecules derived from the gene to be targeted that do not share off-targets. These "clean" RNAi constructs can be used to investigate the consequences of downregulation of a specific gene in the absence of predicted off-target effects. Any phenotypical observations that occur after using both RNAi molecules in separate experiments are then expected to originate solely from its intended target and not from any off-targets for which each RNAi molecule has its own unique set.

Performing a controlled RNAi experiment to model the human disease PANK in *Drosophila melanogaster*

In chapter 4, the new tools as described in the previous chapters were used to design "clean" RNAi constructs used in experiments to study the consequences of downregulation of pantothenate kinase, a gene associated with the human neurodegenerative disease PKAN (Pantothenate kinase-associated neurodegeneration). The disease is characterized by an early onset during childhood and the symptoms are: brain abnormalities, locomotion dysfunction and early death [39]. PKAN is caused by mutations in the pantothenate kinase 2 gene (PANK2), which is the first enzyme in the pathway responsible for the conversion and synthesis of Coenzyme A (CoA) from its precursor vitamin B12 [40]. CoA is an essential metabolic cofactor for many biochemical reactions such as the citric acid cycle and fatty acid oxidation.

A PKAN mouse model was created in which the mouse PANK2 gene was disrupted [41]. Unfortunately this mouse model lacks the neurodegenerative characteristics of PKAN, most likely because mouse PANK2 is in contrast to human PANK2 not localized in mitochondria [42, 43]. These data indicate that using the mouse model is not appropriate to understand the underlying mechanisms of PKAN. However, in *Drosophila melanogaster*, mutations in fumble (the *Drosophila* PANK2 orthologue, further referred to as dPANK/fbl) induce neurodegeneration, locomotion dysfunction and an early death [44]. These data indicate that the *Drosophila* model is appropriate to perform PKAN-related research. A cell line derived from a primary culture of late stage (20-24 hours old) *Drosophila* embryos (S2-cells) was used to create a cell model for PKAN with the use of RNAi. Using the newly developed algorithm as described in chapter 2, a dsRNA construct has been designed devoid of most of the predicted potential off-targets. Western blot analysis revealed a highly efficient knock-down of

dPANK/fbl. The S2 cell model in combination with RNAi technology provided us with a tool to analyse biochemical abnormalities as a result of impaired dPANK/fbl function. In addition different compounds were tested for their rescuing potential towards the dPANK/fbl-depleted phenotype. Chapter 4 describes the observations and the effects of different compounds based on this S2 model. Findings in the S2 cells were further investigated in the Drosophila mutants and appeared to be highly similar.

miRNAs and identifying their targets using RNAi-Select in Hodgkin Lymphoma

In chapter 5, a new bioinformatics tool related to the previous tools was developed to understand the human disease Hodgkin Lymphoma. Hodgkin Lymphoma is a cancerous disease, originating from irrepressible dividing white blood cells (lymphocytes) and characterized by an aberrant miRNA profile [45].

RNA regulation is a natural phenomenon and not restricted to exogenously delivered siRNAs. In fact, endogenous miRNAs are vital during cellular development [46, 47]. In addition, miRNA malfunction has been associated with various pathological conditions including cancer [48-51] due to regulatory imbalances. Studying miRNAs and their targets is a key factor towards understanding the pathology of many cancerous diseases. Unfortunately proper tools allowing genome-wide identification of miRNA targets are lacking. In part because of technical difficulties to detect miRNA changes in biological material, but also due to the nature of miRNAs which allow a much shorter length of sequence homology to be active as compared to siRNAs. A sequence as short as 6-7 nucleotides (the so-called seed region) has proven to induce regulatory effects [52], this is a remarkable finding because any random 6 nucleotide sequence is expected to be found tens of thousands of times in a genome with the size of *Drosophila* (see formula 1 in supplementary file Chapter 3). Obviously, miRNAs or their possible targets must have some other properties that allow for a more specific regulatory function, like for example spatial or temporal expression differences. Also the position of the homologous site on the target mRNA and enriched regions that contain multiple copies of a certain seed region may play a role in determining the real targets. A bio-informatics challenge is to find/predict the natural targets of miRNAs, because based on sequence

homology (or similarity) there are numerous candidates and numerous false positive hits.

Chapter 5 describes an in-depth analysis to find miRNA targets in human cells. By combining bioinformatics and 'wet' experiments, many genes were identified in which the regulation show a strong correlation with the expression of the miRNAs *miR-16, miR-21, miR-24 and miR-155* that are involved in Hodgkin Lymphoma [45]. Research in this field is important as regulatory malfunction is characteristic for cancerous diseases in general and miRNAs are being discovered for playing a central role in them [53].

References

1. Bass BL. Double-Stranded RNA as a Template for Gene Silencing, Cell 2000;101:235-238.
2. Ketting RF, Fischer SEJ, Bernstein E et al. Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans, Genes & Development 2001;15:2654-2659.
3. Siomi H, Siomi MC. On the road to reading the RNA-interference code, Nature 2009;457:396-404.
4. Wei J-X, Yang J, Sun J-F et al. Both Strands of siRNA Have Potential to Guide Posttranscriptional Gene Silencing in Mammalian Cells, PLoS ONE 2009;4:e5382.
5. Sano M, Sierant M, Miyagishi M et al. Effect of asymmetric terminal structures of short RNA duplexes on the RNA interference activity and strand selection, Nucleic Acids Research 2008;36:5812-5821.
6. Fagegaltier D, Bougé A-L, Berry B et al. The endogenous siRNA pathway is involved in heterochromatin formation in Drosophila, Proceedings of the National Academy of Sciences 2009;106:21258-21263.
7. Iorio MV, Piovan C, Croce CM. Interplay between microRNAs and the epigenetic machinery: An intricate network, Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 2010;Epub ahead of print.
8. Gregory RI, Yan K-p, Amuthan G et al. The Microprocessor complex mediates the genesis of microRNAs, Nature 2004;432:235-240.
9. Lin SL, Miller JD, Ying SY. Intronic MicroRNA (miRNA), J Biomed Biotechnol 2006;2006:26818.
10. Didiano D, Hobert O. Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions, Nat Struct Mol Biol 2006;advanced online publication:849-851.

11. Kornberg RD. The molecular basis of eukaryotic transcription, Proceedings of the National Academy of Sciences 2007;104:12955-12961.
12. Alexopoulou L, Holt AC, Medzhitov R et al. Recognition of double-stranded RNA and activation of NF-[kappa]B by Toll-like receptor 3, Nature 2001;413:732-738.
13. Jackson A, Bartz S, Schelter J et al. Expression profiling reveals off-target gene regulation by RNAi, Nature biotechnology 2003;21:635-637.
14. Scacheri PC, Rozenblatt-Rosen O, Caplen NJ et al. Short interfering RNAs can induce unexpected and divergent changes in the levels of untargeted proteins in mammalian cells, Proc Natl Acad Sci U S A 2004;101:1892-1897.
15. Lin X, Ruan X, Anderson MG et al. (2005), 'siRNA-mediated off-target gene silencing triggered by a 7 nt complementation', pp. 4527-4535.
16. Qiu S, Adema C, Lane T. A computational study of off-target effects of RNA interference, Nucleic Acids Res 2005;33:1834-1847.
17. Birmingham A, Anderson E, Reynolds A et al. 3' UTR seed matches, but not overall identity, are associated with RNAi off-targets, Nature methods 2006;3:199-204.
18. Jackson A, Burchard J, Schelter J et al. Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity, RNA 2006;12:1179-1187.
19. Kulkarni M, Booker M, Silver S et al. Evidence of off-target effects associated with long dsRNAs in Drosophila melanogaster cell-based assays, Nat Methods 2006;3:833-838.
20. Ma Y, Creanga A, Lum L et al. Prevalence of off-target effects in Drosophila RNA interference screens, Nature 2006;443:359-363.
21. Moffat J, Reiling JH, Sabatini DM. Off-target effects associated with long dsRNAs in Drosophila RNAi screens, Trends in pharmacological sciences 2007;28:149-151.
22. Saxena S, Jonsson Z, Dutta A. Small RNAs with imperfect match to endogenous mRNA repress translation. Implications for off-target activity of small inhibitory RNA in mammalian cells, The Journal of biological chemistry 2003;278:44312-44319.
23. Pal-Bhadra M, Leibovitch B, Gandhi S et al. Heterochromatic Silencing and HP1 Localization in Drosophila Are Dependent on the RNAi Machinery, Science 2004;303:669-672.
24. Verdel A, Jia S, Gerber S et al. RNAi-Mediated Targeting of Heterochromatin by the RITS Complex, Science 2004;303:672-676.
25. Matzke MA, Birchler JA. RNAi-mediated pathways in the nucleus, Nat Rev Genet 2005;6:24-35.
26. Robb GB, Brown KM, Khurana J et al. Specific and potent RNAi in the nucleus of human cells, Nat Struct Mol Biol 2005;12:133-137.

27. Bosher J, Dufourcq P, Sookhareea S et al. RNA Interference Can Target Pre-mRNA: Consequences for Gene Expression in a Caenorhabditis elegans Operon, Genetics 1999;153:1245-1256.
28. Naito Y, Yamada T, Ui-Tei K et al. siDirect: highly effective, target-specific siRNA design software for mammalian RNA interference, Nucleic Acids Res 2004;32:W124-129.
29. Naito Y, Yamada T, Matsumiya T et al. dsCheck: highly sensitive off-target search software for double-stranded RNA-mediated RNA interference, Nucleic Acids Res 2005;33:W589-591.
30. Smith TF, Waterman MS. Identification of common molecular subsequences, J Mol Biol 1981;147:195-197.
31. Wirawan A, Kwoh C, Hieu N et al. CBESW: Sequence Alignment on the Playstation 3, BMC Bioinformatics 2008;9:377.
32. Janaki C, Joshi RR. Accelerating comparative genomics using parallel computing, In Silico Biol 2003;3:429-440.
33. Schadt EE, Linderman MD, Sorenson J et al. Computational solutions to large-scale data management and analysis, Nat Rev Genet;11:647-657.
34. Zeng Y, Cullen BR. RNA interference in human cells is restricted to the cytoplasm, RNA 2002;8:855-860.
35. Bond D, Foley E. A Quantitative RNAi Screen for JNK Modifiers Identifies Pvr as a Novel Regulator of *Drosophila* Immune Signaling, PLoS Pathog 2009;5:e1000655.
36. Gobert V, Osman D, Bras S et al. A Genome-Wide RNA Interference Screen Identifies a Differential Role of the Mediator CDK8 Module Subunits for GATA/ RUNX-Activated Transcription in Drosophila, Mol. Cell. Biol.;30:2837-2848.
37. Liu T, Sims D, Baum B. Parallel RNAi screens across different cell lines identify generic and cell type-specific regulators of actin organization and cell morphology, Genome Biology 2009;10:R26.
38. Fedorov Y, Anderson EM, Birmingham A et al. Off-target effects by siRNA can induce toxic phenotype, RNA 2006;12:1188-1196.
39. Gregory A, Polster BJ, Hayflick SJ. Clinical and genetic delineation of neurodegeneration with brain iron accumulation, Journal of Medical Genetics 2009;46:73-80.
40. Zhou B, Westaway SK, Levinson B et al. A novel pantothenate kinase gene (PANK2) is defective in Hallervorden-Spatz syndrome, Nat Genet 2001;28:345-349.
41. Kuo Y-M, Duncan JL, Westaway SK et al. Deficiency of pantothenate kinase 2 (Pank2) in mice leads to retinal degeneration and azoospermia, Human Molecular Genetics 2005;14:49-57.

42. Hörtnagel K, Prokisch H, Meitinger T. An isoform of hPANK2, deficient in pantothenate kinase-associated neurodegeneration, localizes to mitochondria, Human Molecular Genetics 2003;12:321-327.
43. Leonardi R, Zhang Y-M, Lykidis A et al. Localization and regulation of mouse pantothenate kinase 2, FEBS Letters 2007;581:4639-4644.
44. Bosveld F, Rana A, van der Wouden PE et al. De novo CoA biosynthesis is required to maintain DNA integrity during development of the Drosophila nervous system, Human Molecular Genetics 2008;17:2058-2069.
45. Gibcus JH, Tan LP, Harms G et al. Hodgkin lymphoma cell lines are characterized by a specific miRNA expression profile, Neoplasia 2009;11:167-176.
46. Hildebrand J, Rutze M, Walz N et al. A Comprehensive Analysis of MicroRNA Expression During Human Keratinocyte Differentiation In Vitro and In Vivo, J Invest Dermatol.
47. Kuipers H, Schnorfeil FM, Brocker T. Differentially expressed microRNAs regulate plasmacytoid vs. conventional dendritic cell development, Molecular Immunology;In Press, Corrected Proof.
48. Cho WC. MicroRNAs as therapeutic targets for lung cancer, Expert Opinion on Therapeutic Targets;14:1005-1008.
49. Weber JA, Baxter DH, Zhang S et al. The MicroRNA Spectrum in 12 Body Fluids, Clin Chem:clinchem.2010.147405.
50. Bhagavathi S, Czader M. MicroRNAs in Benign and Malignant Hematopoiesis, Archives of Pathology & Laboratory Medicine;134:1276-1281.
51. Tan LP, Seinen E, Duns G et al. A high throughput experimental approach to identify miRNA targets in human cells, Nucleic Acids Res 2009.
52. Brennecke J, Stark A, Russell RB et al. Principles of microRNA-target recognition, PLoS Biol 2005;3:e85.
53. Sharma S, Kelly TK, Jones PA. Epigenetics in cancer, Carcinogenesis;31:27-36.

Chapter 2

A Genome-Wide analysis of the specificity of RNAi in Drosophila melanogaster using the novel tool RNAiSelect

Erwin Seinen, Ritsert C. Jansen, Ody C.M. Sibon

An adapted version is currently under revision for publication in Briefings in Functional Genomics (2011)

Author Summary

Genes can be silenced with short-interfering-RNA molecules (siRNA). siRNAs are widely used to identify gene functions and have high potential for therapeutic treatments. It is critical that the siRNA specifically targets the expression of the gene of interest but has no off-target effects on other genes. Although siRNAs were initially considered to be exclusively active on mature mRNAs in the cytoplasm, additional studies have shown that siRNAs are present in the nucleus as well. In this study we investigated whether nuclear intron-containing premature mRNAs should be considered in off-target profiling. By using *Drosophila melanogaster* micro-array data we indeed show a significant off-target occurrence on sequences with homology not only to exonic but also to intronic sequences. Therefore, accurately predicting off-targets at a genome-wide level is important, but validated algorithms that seek beyond the mature mRNA sequences were not available. We designed the novel tool *RNAiSelect* to make comprehensive off-target profiling, based on sequence homology throughout the genome available to the public. With this tool we profiled 1.5 million RNAi sequences derived from all *D. melanogaster* genes.

Introduction

Off-target effects are caused by unintended cross-hybridization between siRNAs and endogenous RNA sequences other than the targeted sequences (1-5). They obscure the functional interpretation of gene silencing experiments (1) and should therefore be avoided as much as possible. Potential hybridizations between siRNAs and mature mRNAs are generally *in silico* analyzed with the user friendly and popular tool BLAST (basic local alignment search tool (6). However, BLAST has insufficient sensitivity for short sequence alignments with partial homology and will likely result in many false negatives. Other more sensitive tools which are also available through a web-interface and specifically designed to find RNAi off-targets (7-9) only apply to mature mRNA sequences, but not to promoter, intron and intergenic sequences. However, studies have shown that siRNAs also act in the nucleus (10-15) where they target promoters (15) and introns (10), while intergenic sequences could also account for (as yet unknown) regulatory functions. Therefore, an accurate whole genome alignment tool to identify RNAi sequences that have the smallest chance of inducing off-targets is imperative, and such a tool was not available to the general public.

In order to fulfill this need, we designed a novel algorithm and tool called RNAiSelect. RNAiSelect rapidly scans the genome for potential partial homology with siRNAs of 21 nucleotides in length. Our tool searches for exact, or nearly exact, homology. It will report homology within a short stretch of contiguous nucleotides with up to 3 mismatches, even if there are evenly distributed mismatches which remain undetected in BLAST. It will identify near exact homologous sequences containing G:U wobble mismatches which exhibit a high binding energy (16,17). It will also allow for single nucleotide polymorphisms, increasing the general applicability of pre-analyzed siRNA sequences in studying multiple genotypes of the same species.

Our tool permits the user to freely and rapidly select the best RNAi constructs possible based on sequence homology and thereby keeping off-targets in general to an absolute minimum.

Results

Experimental validation of RNAiSelect

Currently no validated alignment tools exist that predict off-targets at the genome-wide level. In order to obtain such an instrument we first developed a tool to search the entire *Drosophila melanogaster* genome (including introns) for exact, or nearly exact, sequence homology for a short stretch of contiguous nucleotides with up to 3 mismatches. This tool enables an accurate prediction of off-target effects (see Methods) for any RNAi sequence of interest. In *D. melanogaster*, dsRNA molecules of 300-800 base pairs are commonly used to induce down-regulation of genes. From a specific dsRNA, several siRNAs will be formed through the endogenous RNA interference (RNAi) machinery and each siRNA in theory has its own set of potential off-targets. We have tested whether these individual siRNAs could have any off-target effects and if so whether introns are also being targeted. First, our tool was used to predict all off-targets of siRNAs derived from specific dsRNA constructs and subsequently biological data were used to validate the predictive capacity of our tool. We analyzed 6 independent dsRNA experiments for which microarray data are publicly available (see Methods). Using previously described criteria (18-20) (see also supplementary Table 1), we extracted all the potent siRNAs from the dsRNA sequences used and searched for homology against the genome for 21 nucleotides (nt) with up to three mismatches. With the use of our tool we found an average of 83 potential off-targets per dsRNA with 0% of them containing zero mismatches, 4-10% containing one or two mismatches, and 90-96% of them containing 3 mismatches (see below for homology searching with 19 nucleotides).

To estimate the number of true off-targets, we first identified a group of potential off-targets using RNAiSelect. Within this group, we calculated the percentage of transcripts that were actually downregulated on the microarrays (further referred to as “the predicted set”) and compared them to the percentage of downregulated transcripts of the total array which represents a random set collected from the same microarray (see Figure 1).
Together these data show that the percentages of downregulation in the predicted sets are 10.48%, 13.38%, 13.73%, 13.71%, 25.10%, and 17.91% enriched compared to a representative random set of the total array

(P<0.009, P<0.015, P<0.023, P<0.005, P<0.008, P<0.002) in studies 1-6 respectively (Figure 1, see Methods). This provides strong evidence that at least 8 (on average 13) out of the predicted 83 off-target genes were true (and unwanted) off-targets of the siRNAs. Currently web-based tools that allow a comparable type of analysis predicted none, or far fewer, of these validated off-targets, even when the most sensitive parameters for these tools were selected (see Supplementary Table 2).

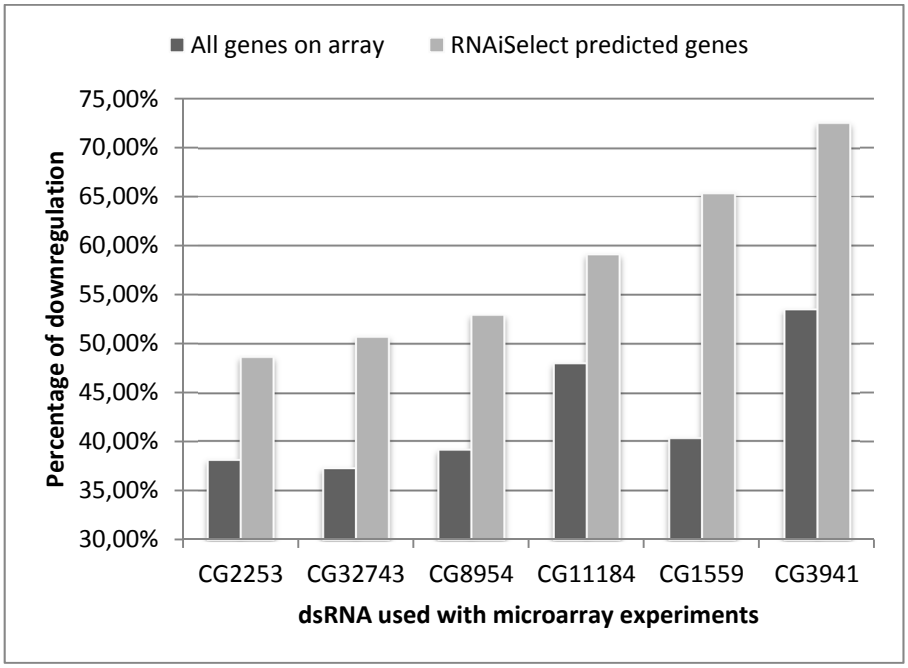


Figure 1. *To validate RNAiSelect, we compared the number of transcripts downregulated on the whole array with the number of predicted off-targets that were actually downregulated. This was performed for six independent experiments. The dark grey bars represent the percentage of downregulated genes of each experiment on the whole array. The light grey bars represent the percentage of predicted off-targets that are downregulated. This analysis shows that the off-target set predicted by RNAiSelect contains an increased fraction of downregulated genes compared to genes randomly selected from the total array (P<0.01 on average).*

Separating exonal from intronal off-targets

RNAiSelect can search for near exact homologies in exons, promoter regions, introns and untranslated regions (UTRs). It can therefore search for

homologies in intergenic sequences that may account for (as yet unknown) regulatory functions. From the potential off-targets predicted by RNAiSelect, we focus on the subset containing exon sequences (set A) and another subset containing intron sequences, including sequences overlapping intron/exon boundaries (set B). The dataset containing the intron sequences (set B) showed the fraction of downregulated genes to be significantly enriched compared to the whole data sets for 3 out of 6 dsRNAs ($P<0.184$, $P<0.026$, $P<0.233$, $P<0.013$, $P<0.1096$, $P< 0.001$ in studies 1-6, respectively in Table 1, see Methods). This finding is consistent with previous findings that there is indeed RNAi activity in the nucleus and specific pre-mRNAs might be exposed to silencing (10,13). This data supports our hypothesis that accurate *in silico* off-target screenings should include exons as well as introns. Examples of predicted off-targets that are strongly downregulated are listed in Supplementary Table 3, many of which are intronic and functionally unrelated to the target gene (Supplementary Table 4). So far, RNAiSelect is the only off-target search tool that allows the user to extend the search to intron-containing pre-mRNA sequences.

	CG2253 (n=142)	**CG32743 (n=73)**	**CG8954 (n=70)**	**CG11184 (n=93)**	**CG1559 (n=26)**	**CG3941 (n=70)**
intron+exon (A+B)	10%	13%	13%	14%	25%	18%
exon only (A)	13%	9%	14%	5%	22%	13%
intron only (B)	8%	19%	11%	20%	30%	32%

Table 1. *The set of predicted potential off-targets (predicted set A+B) was split into one set containing exon sequences (predicted set A) and another set containing intron sequences and sequences overlapping intron/exon boundaries (predicted set B). Enrichment of downregulated genes within the predicted off-targets compared to the microarray background is presented for each set. All the analyzed microarrays showed an increased downregulated fraction in the intronic set (including the boundaries) compared to the background, three of which were statistically significant (see Methods). The number of predicted off-targets per dsRNAi construct for the six independent experiments is indicated with n.*

Allowing for even weaker partial homology

Our analysis shows that homology for 21 nucleotides containing up to 3 mismatches can cause significant off-target effects. Since it has been shown that a homology for 17 contiguous nucleotides can cause off-target effects

(1), it could be important to search for partial homology for 19 contiguous nucleotides with up to 3 mismatches, despite the fact that the set of potential off-targets will most likely contain many more false positives. Performing the analysis using these parameters (including exons, introns and intergenic regions) revealed 5,669, 3,956, 4,255, 6,125, 1,435 and 2,456 potential off-targets for the total sum of the siRNAs that may derive from the dsRNAs of CG2253, CG32743, CG8954, CG11184, CG1559 and CG3941, respectively (Figure 2). It is likely that this set of potential off-targets will contain many false positives, so that there is no longer a significant enrichment of downregulation in the predicted off-target set. However, to reduce the risks in RNAi experiments as much as possible, one can use our tool to select those siRNA constructs that have the lowest number of predicted off-targets (see Figure 2). RNAiSelect can instantly identify these siRNA constructs as well as their potential predicted off-targets, allowing users to perform experiments with the best possible RNAi constructs available based on sequence homology.

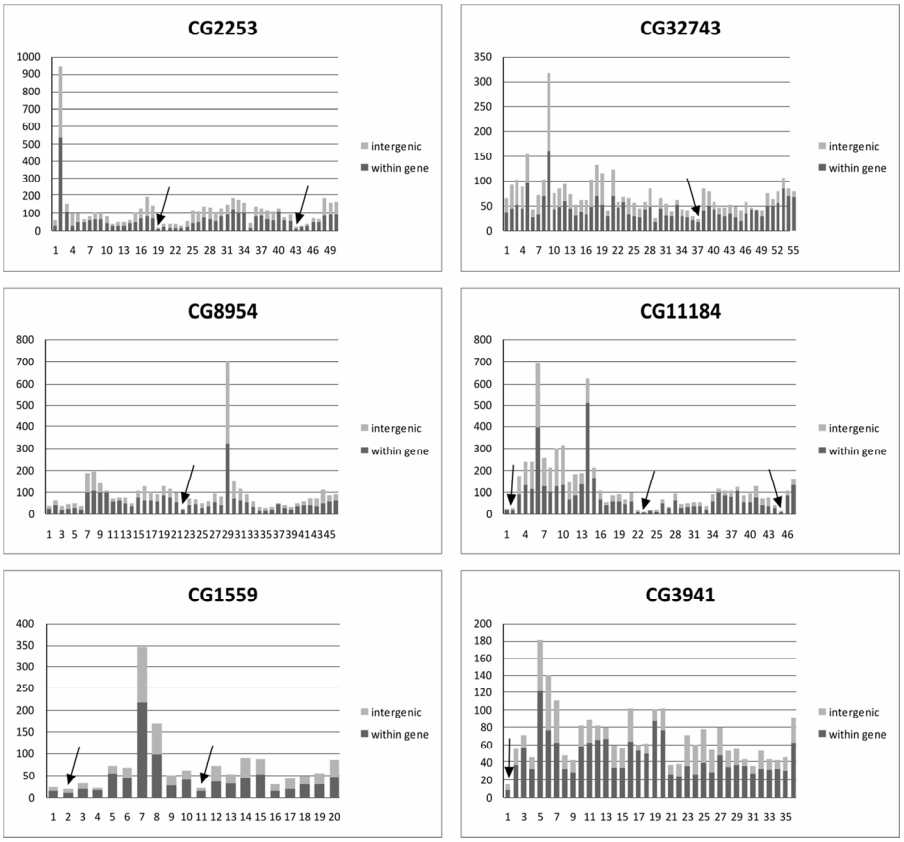


Figure 2. *Predicted off-targets identified by RNAiSelect for the 6 analyzed genes. On the x-axis, all the siRNAs derived from the original dsRNA strand are represented as numbers. On the y-axis, the number of identified off-targets is given. Arrows indicate siRNAs with a low number of predicted off-targets, classifying these rare specific siRNAs as highly specific.*

Analysis of candidate off-targets for all *D. melanogaster* genes

Our analysis of the 6 genes (CG2253, CG32743, CG8954, CG11184, CG1559 and CG3941, shown in Figure 2), reveals a large number of possible off-targets per gene. Next we investigated whether these findings were representative for the complete genome. We used RNAiSelect to scan for all possible siRNAs derived from all annotated *D. melanogaster* genes. Using the same criteria as before (see supplementary Table 1), we extracted 1.5 million candidate siRNA sequences for *D. melanogaster*. On average, we found 121 potential off-targets per siRNA when searching for homology

with a minimum of 19 consecutive nucleotides and 3 mismatches, respectively (median 76). These data show that the dsRNA sequences as analyzed in Figure 2 are not an exception with regard to their large number of off-targets.

MicroRNA effects

It is known that siRNAs can act as microRNAs (miRNAs) based on homology with short seeds of no more than 6 nucleotides on the 3' UTR sequences of the mRNA (21). To find these possible interactions, RNAiSelect offers the possibility to perform a 3' UTR seed pairing analysis using known seed pairing rules (22). Due to the short sequence length (6-nt), there is only a limited number of sequences possible ($4^{\wedge 6}$=4.096). Therefore, if a random genome with the size of D. melanogaster is considered (140 Mb; Flybase R5.20), than any particular seed sequence of 6-nt would theoretically be present a little over 34.000 times throughout the genome (not counting overlapping sequences). Considering 14.419 3' UTR sequences with an average size of 372 nt (Flybase, R5.20), almost 1 out of 10 genes will have any random 6-nt sequence present within its 3' UTR. To show more relevant data that surpasses the background, we only considered the less frequent multiple seed hits within the same 3' UTR as possible miRNA targets. This approach has been demonstrated to decrease the false positive rates (21). In addition, we have recently described a method to *in silico* validate miRNA targets that were identified by *in vivo* experiments (23). Both strategies (21, 23) are employed in the miRNA seed scanning tool within RNAiSelect. This results in the identification of genes that are potentially regulated by siRNAs acting in a miRNA-like manner. From these lists the user can decide either not to use the siRNA or to keep the candidate off-targets in mind while interpreting the results of the experiment.

Generating a resource to select highly specific siRNAs

When selecting the most specific siRNA construct it is not only important to select a sequence with a low number of predicted off-targets (as shown in Figure 2), but also the type of off-target sequences may be relevant. In order to analyze the bulk of data for the type of off-targets at the genome-wide

level we classified the off-targets of an siRNA using the criteria described in Table 2. For this qualification the lowest score (1) was given to an off-target that contains 0 mismatches; score 2 represented an off-target containing 1 G:U wobble mismatch; score 3 represented an off-target containing 1 true mismatch etc. and the highest score was given to an off-target containing 3 true (no G:U wobble) mismatches. The average classification score of all the off-targets of the 1.5 million siRNAs aligned against the genome was calculated to be 7.8. This indicates that, on average, any off-target has 3 mismatches (consisting of 3 G:U wobble mismatches or consisting of 2 G:U wobble mismatches and 1 true mismatch). To further analyze the nature of off-targets, the 1.5 million siRNAs were therefore given the score of their lowest class. With these criteria, the 1.5 million siRNAs have a calculated average of 4.2. This implies that, on average, any siRNA has at least one off-target with 2 G:U wobble mismatches. Together, our genome-wide analysis shows that a randomly selected siRNA construct could have a large number of predicted off-targets with 2 G:U wobble mismatches. The results of these analyses are available as open source (https://sourceforge.net/projects/rnaiselect/files/) and can be used to design future RNAi studies by filtering out these cross-reacting RNAi molecules and instead actively selecting specific criteria for the siRNAs present to downregulate the gene of interest. In case future research reveals specific criteria to allow accurate prediction of the activity of siRNA sequences, the classification table can be adjusted accordingly. Due to the fast speed and internal design of RNAiSelect, it is possible to accommodate a webserver to provide real-time online accessibility.

Class	Description
1	0 mismatches
2	1 G:U wobble mismatch
3	1 true mismatch
4	2 G:U mismatches
5	1 G:U and 1 true mismatch
6	2 true mismatches
7	3 G:U wobble mismatches
8	2 G:U mismatch + 1 true mismatches
9	1 G:U mismatch + 2 true mismatches
10	3 true mismatches

Table 2. *Classification of off-targets, with class 1 having a perfect off-target match and higher numbers (class 2, class 3, etc.) representing classes with increasing numbers of mismatches. G:U wobble mismatches are considered to be more stable than other mismatches and are therefore more represented in the lower classes.*

Computational challenges

The genome-wide alignment of 1.5 million potential siRNA sequences, as performed in this study, is a major computational task. Standard algorithms like BLAST and Smith-Waterman (SW) (23) for local sequence alignment are either not sensitive enough (BLAST) or computationally too intensive (SW). We have therefore developed the RNAiSelect algorithm based on using simple look-up tables. The look-up table contains the exact *D. melanogaster* genomic location for every short sequence of 9 nucleotides. Multiple look-ups of consecutive 9 nucleotide subsequences of a siRNA can find off-target sequences without actual alignment of the siRNA to the genome but by using integer calculations (see Methods). Such look-up tables can easily be constructed, so that the same exercise we have done for *D. melanogaster* can be repeated for other organisms: we anticipate that a similar level of improvement of siRNA specificity can be achieved. RNAiSelect is up to 10 times quicker than SW (24) and equally sensitive.

Statistical analysis

Table 3 and Table 4 show a detailed chi-square analysis of the 6 individual dsRNA experiments from the off-target data predicted by RNAiSelect, with the exception of CG1559 in Table 4 where a two-tailed binomial test was used due to a low transcript number. Table 3 confirms that when considering both introns and exons, the six different dsRNAs show a significant number of off-targets by comparing the observed number of downregulated transcripts with the microarray background ($\alpha = 0.05$). Table 4 shows that when only introns are considered, 3 out of 6 analyses still show a significant number of downregulated transcripts due to off-targets ($\alpha = 0.05$). From this data we concluded that intron-based off-targets were indeed occurring.

Gene	**CG2253**		**CG32743**		**CG8954**		**CG11184**		**CG1559**		**CG3941**	
H0: π =	0.38		0.37		0.39		0.48		0.4		0.53	
Expression	+	-	+	-	+	-	+	-	+	-	+	-
Expected	88.0	54.0	46.0	27.0	43.3	27.7	55.6	51.4	15.6	10.4	32.9	37.1
Observed	73	69	36	37	34	37	41	66	9	17	20	50
Chi-Square	6.761		5.865		5.132		8.025		6.981		9.543	
P-value	0.009		0.015		0.023		0.005		0.008		0.002	

Table 3. *Statistical analysis for both intron and exon data in the microarray experiments For each dsRNA, we analyzed the off-targets predicted by RNAiSelect. The second row presents the percentage of genes that are downregulated on the complete microarray. The number of off-targets as predicted by RNAiSelect are divided in upregulated and downregulated genes (expected values; fourth row) based on the microarray background. The fifth row presents the actual number of upregulated (+) and downregulated (-) genes within the set of off-targets predicted by RNAiSelect (observed values). Chi-square and binomial analysis of these data shows that for every experiment the predicted set of off-targets contains a significant larger fraction of downregulated genes as compared to the complete microarray (H0; second row).*

Gene	**CG2253**		**CG32743**		**CG8954**		**CG11184**		**CG1559**		**CG3941**	
H0: π =	0.38		0.37		0.39		0.48		0.4		0.53	
Regulation	+	-	+	-	+	-	+	-	+	-	+	-
Expected	43.3	26.7	18.8	11.2	17	11	32.3	29.7	6	4	9.4	10.6
Observed	38	32	13	17	14	14	20	42	3	7	3	17
Chi- Square	1.768		4.978		1.424		6.156		[binomial]		11.594	
P-value	0.184		0.026		0.233		0.013		0.1096		0.001	

Table 4*. Statistical analysis for intron data in the microarray experiments As in Table 3, we have analyzed the off-targets predicted by RNAiSelect for each dsRNA, except we now filtered for intron*

targeted regions. For each dsRNA, we further analyzed the by RNAiSelect predicted off-targets. The second row shows the percentage of genes that are downregulated on the complete microarray. The third row presents the number of off-targets as predicted by RNAiSelect. This number is divided in upregulated and downregulated genes (expected values; fourth row) based on the microarray background. The fifth row presents the actual number of upregulated (+) and downregulated (-) genes within set of off-targets predicted by RNAiSelect (observed values). Chi-square and binomial analysis of these data shows that for every experiment the predicted set of off-targets contains a significant larger fraction of downregulated genes as compared to the complete microarray (H0; second row).

Discussion

There are 2 explanations why a non-targeted transcript is downregulated as a results of an RNAi experiment: (i) The transcript is a true off-target of the used RNAi constructs, (ii) The downregulated on-target gene triggers a cascade of regulatory effects which result in down regulation of seemingly unrelated gene products. Explanation (ii) complicates the validation of off-target prediction algorithms and because of this possibility a significantly down regulated transcript does not necessarily represent an off-target effect. Consequently, to enable a proper validation, (a) the 'background noise' has to be corrected for and (b) the dataset to work with must be large enough to enable testing whether there is a significant correlation between downregulated transcripts and identified off-targets. For an optimal correction of the background noise, we compared the number of down regulated genes in the by RNAiSelect predicted set with a randomized group from the same microarray data. Any background noise due to technical limitations or due to specific on-target effects are present in both sets and is corrected for by this approach.
For our analysis we did not use a specific threshold or cutoff value but we divided the transcripts in 2 groups: downregulated and not downregulated. We used this approach for the following reasons. (i) Our aim was not to identify individual transcripts to be downregulated but instead we were interested in a general trend. (ii) Previously, it has been demonstrated that RNAi can induce off-target effects resulting in less than 2-fold reduced expression (25), while still inducing strong protein reduction and subsequently biological effects. (iii) By using no threshold, the analysis contains many more transcripts which enhanced the sensitivity to a great extent. Moreover this allows a proper statistical analysis, which would not be possible when small groups were used. In addition, by maximizing the sensitivity, small significant expression changes that might have real

biological effects are not overlooked. With these considerations, we validated our methods on six experimental data sets from *D. melanogaster*. *In silico* we predicted the potential off-targets of specific double-stranded RNAs (dsRNAs) and empirically show that predicted off-target genes were significantly more frequently silenced than other genes. Moreover, we show that intron containing off-target effects and homologies up to 3 mismatches should not be ignored.
It is to be expected that the off-target RNAi activity is reversely proportional with the number of mismatches. Unfortunately we were unable to make any statistical distinction between the number of mismatches in our dataset, for the reason that off-targets with < 3 mismatches are relatively rare in comparison to 3 or more mismatches. However, because 90-96% of the predicted off-targets do contain 3 mismatches and therefore account for the majority in the significant enrichment during our validation, our data demonstrate that this type of off-target should not be ignored and that partial homology searches are indeed necessary.

Due to the less stringent homology requirements, RNAiSelect will most likely over-predict the number of off-targets. However, this approach is valid, because RNAiSelect is used to find RNAi molecules that have a low number of predicted off-targets to minimize potential side-effects. These sequences with the least amount of off-targets can then be used for knock-down experiments. From that view, over-predicting off-targets is far better than being unaware of possible off-targets, while at least suspicious candidates (like the examples listed in Supplementary Table 3) can now be identified.
Fortunately, in our genome-wide analysis of all potential siRNA sequences, we have found that, when selecting for the most specific siRNAs, the majority of genes have potent siRNAs with 24 or fewer predicted off-targets (arrows in Figure 2). These selected siRNAs have 80% fewer predicted off-targets than the average 121 off-targets generally found. In addition to select for specific siRNAs, the user can make an inventory of possible off-targets when using a particular dsRNAs. This will allow to evaluate existing expression profiles derived from experiments using RNAi technology, e.g. to identify false-positive results caused by homologous induced off-target effects of the used dsRNA sequences. When performing a genome-wide screen using dsRNA molecules, RNAiSelect might also be useful to assess the positive hits within this screen and assist in the decision which of them are most promising to proceed with.

In addition to a real-time analysis tool, we supply a comprehensive database containing 1.5 million pre-analyzed siRNAs covering the whole genome of *D. melanogaster*. RNAiSelect uses this database to allow the generation of a detailed report containing the number and type of mismatches which assists in rapidly selecting specific siRNAs while keeping potential off-targets to a minimum. These siRNAs can then be used instead of the less specific and more generally used dsRNAs (26). Selecting for the most specific siRNAs will be even more important when cocktails of siRNAs are used to downregulate multiple gene products as will be of value for complex traits studies (27).
In conclusion, our tool identifies many validated off-targets, which results in simple and rapid identification of those rare siRNAs with few potential off-target effects. Information on the most selective siRNAs for any individual gene is generated for the users of RNAiSelect, allowing them to choose those siRNAs with the smallest likelihood for off-target effects. Although our approach does not give detailed insights in why specific RNAi constructs are effective, our tool permits the user to work with the best RNAi constructs possible based on sequence homology and thereby keeping off-targets in general to an absolute minimum. Considering the conservation of RNAi mechanisms across species, our findings in *D. melanogaster* will also be of interest for research based on other model systems in which RNAi technology is applied.

Methods

Computer hardware and software

Genomic data (build 45-43b) were downloaded from the Ensembl website (www.ensembl.org). The data from Ensemble and its derived seed tables were processed, stored and indexed in a MySQL database, version 5.0, running on top of Ubuntu 6.06. RNAiSelect was written in C#.NET and can run as a standalone command line executable or within a Microsoft Internet Information Server environment connected to the MySQL database server. Both database and application were hosted on a single system with dual XEON 5140 2.33 GHz processors and 16 GB of 667 ECC memory.

RNAiSelect algorithm

The RNAiSelect algorithm was specifically designed for finding relationships between short nucleotide sequences. It has a high performance and usability for any short-sequence study, including siRNA (off-)targets or miRNA docking sites. The algorithm is based on the following assumption:
"An example sequence TTTTAATTTGGGCCCGGG consists of 18 nucleotides and may be split into two 9-nt child sequences; TTTTAATTT and GGGCCCGGG. By plain observation, we know that the sequence GGGCCCGGG is exactly 9-nt separated from TTTTAATTT in the original sequence."
For the RNAiSelect algorithm to work, we first wrote a program that generates a seed table which holds the exact *D. melanogaster* genomic location(s) for every possible 9-nt sequence ($4^{\wedge 9}$, or 262.144 sequences). Generating such an index is a general strategy used by many algorithms to rapidly look-up any sequence of fixed length for its positions in the genome. Our algorithm however uses a novel method to calculate the positional relationship between indexed seeds, instead of performing string-to-string comparisons for every nucleotide after a hit has been found. In other words, by searching 9-nt subsequences of the whole query sequence for consecutive matches of locations, it will find hits larger than 9nt without performing actual DNA comparisons. This following example, in layman code, shows how to find an 18-nt sequence in the genome by first splitting the sequence into its two 9-nt subsequences and comparing these sequences with the available index table with a word size of 9.

1	SPLIT QUERY SEQUENCE(18 nt) INTO *dnacode_left(9 nt)* AND *dnacode_right(9 nt)*
2	EXTRACT LOCATIONS FROM *index_table* FOR dnacode_left AND STORE IN *seedtable_left*
3	EXTRACT LOCATIONS FROM *index_table* FOR dnacode_ *right* AND STORE IN *seedtable_* right
4	SELECT ALL HITS WHERE (LOCATIONS *seedtable_left + 9) EQUALS (LOCATIONS seedtable_right)*

This example merely demonstrates how to find an exact 18-nt hit not allowing any mismatches. However, users can allow mismatches by expanding the seed searches by variations in such a way that all possible

combinations will be found. Supplementary Figure 1 shows how mismatches may be distributed on a single 18-nt sequence. We thus included variations of the 9-nt sub-sequences and then compared the distance relationship between the original locations of the seed hits, which has to be exactly 9. This may considerably increase the number of seed searches, but because these are relatively cheap in terms of processing time, the overall performance is very high while it guarantees that every possible alignment is evaluated.

Validation by microarray analysis

Microarray data were obtained via the EMBL-EBI online repository (http://www.ebi.ac.uk/) (28). We have used the microarray data with the IDs MEXP-202 and E-GEOD-2623. The downloaded raw CEL-data were imported into ArrayAssist 5.5.1 and PLIER normalized. Because our analysis requires all information available at the micro-arrays, we used an approach that allows the evaluation of transcript levels within a large group. First, all transcripts of the whole micro array derived from the dsRNA experiments (the primers used to construct the dsRNA sequences are listed in Supplementary Table 5) were divided in two groups: one group representing all upregulated transcripts and another group representing all downregulated transcripts as compared to the control array. N.B for this specification of groups no cut-off values were used. Although this analysis is not appropriate for single probe analysis, this approach does make it possible to gather sufficient information to identify a general trend within the chosen groups as compared to the background. All ratio comparisons were subjected to chi-square or binomial analysis (see Table 3 and 4) to find significant trends. For all 6 experiments the ratio of down- versus upregulated transcripts was defined and referred to as the background ratio (presented in Figure 1). RNAiSelect was used to define the predicted set of off-targets and within these sets, the ratio of down- versus upregulated transcripts was determined. This ratio was compared with a randomized group representing the background ratio (presented in Figure 1).

Acknowledgements

We thank Gerald de Haan and Lenoid Bystrykh for critical reading of the manuscript. We thank Norbert Perrimon, Matthew Brooker and Bernard Mathey-Prevot for essential advice and stimulating discussions. We also thank Hans Burgerhof for his detailed statistical analysis.

Financial Disclosure

This work was supported by a VIDI grant from the Netherlands Organization for Scientific Research (NWO; 971-36-400) to O.C.M.S. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Kulkarni, M., Booker, M., Silver, S., Friedman, A., Hong, P., Perrimon, N. and Mathey-Prevot, B. (2006) Evidence of off-target effects associated with long dsRNAs in Drosophila melanogaster cell-based assays. *Nat Methods*, **3**, 833-838.
2. Moffat, J., Reiling, J.H. and Sabatini, D.M. (2007) Off-target effects associated with long dsRNAs in Drosophila RNAi screens. *Trends in pharmacological sciences*, **28**, 149-151.
3. Fedorov, Y., Anderson, E.M., Birmingham, A., Reynolds, A., Karpilow, J., Robinson, K., Leake, D., Marshall, W.S. and Khvorova, A. (2006) Off-target effects by siRNA can induce toxic phenotype. *RNA*, **12**, 1188-1196.
4. Ma, Y., Creanga, A., Lum, L. and Beachy, P. (2006) Prevalence of off-target effects in Drosophila RNA interference screens. *Nature*, **443**, 359-363.
5. Jackson, A., Burchard, J., Schelter, J., Chau, B., Cleary, M., Lim, L. and Linsley, P. (2006) Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity. *RNA*, **12**, 1179-1187.
6. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403-410.
7. Iyer, S., Deutsch, K., Yan, X. and Lin, B. (2007) Batch RNAi selector: a standalone program to predict specific siRNA candidates in batches with enhanced sensitivity. *Comput Methods Programs Biomed*, **85**, 203-209.
8. Naito, Y., Yamada, T., Matsumiya, T., Ui-Tei, K., Saigo, K. and Morishita, S. (2005) dsCheck: highly sensitive off-target search software for double-stranded RNA-mediated RNA interference. *Nucleic Acids Res*, **33**, W589-591.
9. Naito, Y., Yamada, T., Ui-Tei, K., Morishita, S. and Saigo, K. (2004) siDirect: highly effective, target-specific siRNA design software for mammalian RNA interference. *Nucleic Acids Res*, **32**, W124-129.
10. Bosher, J., Dufourcq, P., Sookhareea, S. and Labouesse, M. (1999) RNA Interference Can Target Pre-mRNA: Consequences for Gene Expression in a Caenorhabditis elegans Operon. *Genetics*, **153**, 1245-1256.
11. Langlois, M.-A., Boniface, C., Wang, G., Alluin, J., Salvaterra, P., Puymirat, J., Rossi, J. and Lee, N. (2005) Cytoplasmic and Nuclear Retained DMPK mRNAs Are Targets for RNA Interference in Myotonic Dystrophy Cells. *J Biol Chem*, **280**, 16949-16954.

12. Matzke, M.A. and Birchler, J.A. (2005) RNAi-mediated pathways in the nucleus. *Nat Rev Genet*, **6**, 24-35.
13. Robb, G.B., Brown, K.M., Khurana, J. and Rana, T.M. (2005) Specific and potent RNAi in the nucleus of human cells. *Nat Struct Mol Biol*, **12**, 133-137.
14. Weinberg, M.S., Barichievy, S., Schaffer, L., Han, J. and Morris, K.V. (2007) An RNA targeted to the HIV-1 LTR promoter modulates indiscriminate off-target gene activation. *Nucleic Acids Res*, **35**, 7303-7312.
15. Morris, K., Simon, W.L.C., Jacobsen, S. and Looney, D. (2004) Small Interfering RNA-Induced Transcriptional Gene Silencing in Human Cells. *Science*, **305**, 1289-1292.
16. Xu, D., Landon, T., Greenbaum, N.L. and Fenley, M.O. (2007) The electrostatic characteristics of G.U wobble base pairs. *Nucleic Acids Res*, **35**, 3836-3847.
17. Holen, T., Moe, S., Sorbo, J., Meza, T., Ottersen, O. and Klungland, A. (2005) Tolerated wobble mutations in siRNAs decrease specificity, but can enhance activity in vivo. *Nucleic Acids Res*, **33**, 4704-4710.
18. Jagla, B., Aulner, N., Kelly, P.D., Song, D., Volchuk, A., Zatorski, A., Shum, D., Mayer, T., De Angelis, D.A., Ouerfelli, O. *et al.* (2005) Sequence characteristics of functional siRNAs. *RNA*, **11**, 864-872.
19. Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W. and Khvorova, A. (2004) Rational siRNA design for RNA interference. *Nature biotechnology.*, **22**, 326-330.
20. Elbashir, S.M., Martinez, J., Patkaniowska, A., Lendeckel, W. and Tuschl, T. (2001) Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate. *EMBO J*, **20**, 6877-6888.
21. Birmingham, A., Anderson, E., Reynolds, A., Ilsley-Tyree, D., Leake, D., Fedorov, Y., Baskerville, S., Maksimova, E., Robinson, K., Karpilow, J. *et al.* (2006) 3' UTR seed matches, but not overall identity, are associated with RNAi off-targets. *Nature methods*, **3**, 199-204.
22. Brennecke, J., Stark, A., Russell, R.B. and Cohen, S.M. (2005) Principles of microRNA-target recognition. *PLoS Biol*, **3**, e85.
23. Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J Mol Biol*, **147**, 195-197.
24. Yamada, T. and Morishita, S. (2005) Accelerated off-target search algorithm for siRNA. *Bioinformatics*, **21**, 1316.

25. Aleman, L., Doench, J. and Sharp, P. (2007) Comparison of siRNA-induced off-target RNA and protein effects. *RNA*, **13**, 385-395.
26. Wakiyama, M., Matsumoto, T. and Yokoyama, S. (2005) Drosophila U6 promoter-driven short hairpin RNAs effectively induce RNA interference in Schneider 2 cells. *Biochem Biophys Res Commun*, **331**, 1163-1170.
27. Jansen, R.C. (2003) Studying complex biological systems using multifactorial perturbation. *Nat Rev Genet*, **4**, 145-151.
28. Brazma, A. and Parkinson, H. (2006) ArrayExpress service for reviewers/editors of DNA microarray papers. *Nat Biotech*, **24**, 1321-1322.

Supporting Information Chapter 2

Supplementary Table 1

Description	Score
30%-52% GC Content	1 point
3 or more A/Us at positions 15-19	1 point per A/U
Tm>20°C	1 point
A at position 19	1 point
A at position 3	1 point
U at position 10	1 point
G/C at position 19	-1 point
G at position 13	-1 point
>4 sequential nucleotide repeat	-9 points
> 4 diplet repeat	-9 points

Scoring scheme used to define most potent siRNAs, based on a summary from several publications. Sequences scoring at least 6 point were considered by RNAiSelect.

References:

- http://www.protocol-online.org/prot/Protocols/Rules-of-siRNA-design-for-RNA-interference--RNAi--3210.html
- Jagla, B., Aulner, N., Kelly, P.D., Song, D., Volchuk, A., Zatorski, A., Shum, D., Mayer, T., De Angelis, D.A., Ouerfelli, O. et al. (2005) Sequence characteristics of functional siRNAs. RNA, 11, 864-872.
- Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W. and Khvorova, A. (2004) Rational siRNA design for RNA interference. Nature biotechnology., 22, 326-330.
- Elbashir, S.M., Martinez, J., Patkaniowska, A., Lendeckel, W. and Tuschl, T. (2001) Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate. EMBO J, 20, 6877-6888.

Supplementary Table 2

Comparison of RNAiSelect with BLAST (most widely accepted web-based tool) and dsCheck/siDirect (most accurate web-based tool).

	RNAiSelect	**BLAST**	**dsCheck / siDirect**
Online available web-application	Yes	Yes	Yes
Speed	Fast	Fast	Fast
Accurate short sequence alignment	Yes	**No**	Yes
Finds 1 (G:U or other) mismatches	Yes	Yes	Yes
Finds 2 (G:U or other) mismatches	Yes	**No**	Yes
Finds 3+ (G:U or other) mismatches	Yes	**No**	**No**
Finds exon or UTR based siRNA off-targets	Yes	Some	Yes
Finds whole genome siRNA off-targets, including introns	Yes	Some	**No**
Finds seed based miRNA off-targets	Yes	**No**	**No**
Designs specific shRNAs to knockdown *D. melanogaster* genes[1]	Yes	**No**	**No**
Average validated off-targets found per dsRNA[2]	13	0	2
Identification of non-overlapping dsRNAs with no shared off-targets	Yes	**No**	**No**

[1]In general, long dsRNAs are used for D. melanogaster RNAi experiments. RNAiSelect is the only tool that allows the design of short, specific 21 bp shRNAs. [2]Average number of verified off-targets (by 6 independent microarray data) with 700 bp dsRNAs against 6 different genes.

Supplementary Table 3

Examples of predicted and downregulated off-targets containing various types of mismatches

21-nt sequence from dsRNA	Targeted gene	Off-targeted gene	Regular mismatches	G:U mismatches	exon/intron	fold downregulation
Q: AGCCGAAGGUGCUGAACAAGU ||||||| |||| ||||| R: GGCCGAAGCUGCUGUACAAGU	CG3941	CG3629	3	0	intron	>50
Q: ACAACGACAACGACAUCGAUA ||||| |||| |||||| | R: ACAACAACAACAACAUCGACA	CG2253	CG4128	3	0	intron	>50
Q: CUUUUCGGCUUUGUUUUGAUU ||||| ||||||:|||| || R: CUUUUUGGCUUUGGUUUGUUU	CG11184	CG4128	2	1	intron	>50
Q: AGCACGAAAUCGAAGAGAAAC |||| ||| ||||||||||| R: AGCAGAAAACCGAAGAGAAAC	CG8954	CG2507	3	0	exon	>50
Q: ACAACGACAACGACAUCGAUA ||||| ||||||||:||| | R: ACAACAACAACGACAGCGACA	CG2253	CG2507	2	1	exon	33
Q: ACAACGACAACGACAUCGAUA ||||||||||| ||| || || R: ACAACGACAACAACAACGUUA	CG2253	CG3315	3	0	exon	33
Q: UCGAGGCCAAACUGAAAAUGA |||| |||||||||| ||| R: CCGAGCCCAAACUGAAACUGA	CG2253	CG9656	3	0	intron	20
Q: ACAUCAUGUUUGCAUUUGUUG ||| ||||||||||| |||: R: ACACCAUGUUUGCAUUCGUUU	CG11184	CG1133	2	1	intron	16
Q: AGAACGCGAUCCACCCAGAAA || |||| |||| ||||||| R: AGGACGCAAUCCUCCCAGAAA	CG32743	CG13185	3	0	exon	12

Q: UUAUCAACCGCAAGUCGUAUC |||| |||| |||||||||| R: UUAUGAACCACAAGUCGUAUA	CG32743	CG7978	3	0	intron	11
Q: CUUUUCGGCUUUGUUUUGAUU |||||||| ||||||||| | R: CUUUUCGGUUUUGUUUUGGCU	CG11184	CG12290	3	0	exon	7
Q: UCGGCCUGAUUGGCUUUAUCA ||||| || |||:|||||||| R: UCGGCUUGUUUGUCUUUAUCA	CG32743	CG12819	2	1	exon	7
Q: UGCAACAACUGCCGCAAAUGG |||||||| |||||| |||| R: UGCAACAAAUGCCGCGAAUGC	CG1559	CG15295	3	0	exon	6
Q: UGCAACAACUGCCGCAAAUGG |||||||| ||| |||||||| R: UGCAACAAGUGCAGCAAAUGG	CG1559	CG32046	2	0	intron	6
Q: ACAUCAAGGCCACCGAGAAGA ||||| |:||| ||||||| R: ACAUCACGUCCACGGAGAAGA	CG2253	CG3234	2	1	exon	6
Q: ACAACGACAACGACAUCGAUA ||| |||| |||||||||:| R: ACAUCGACAUCGACAUCGAGA	CG2253	CG32130	2	1	Intron	6
Q: CUGCGUCUGUCCAAGAUCAUC ||| ||||||||||||| R: GUGCCUCUGUCCAAGAUCAUA	CG32743	CG4678	3	0	exon	6
Q: AUCUGCGUCUGUCCAAGAUCA ||||| || |||||||| | R: AUCUGCCUCCGUCCAAGAUGA	CG32743	CG3359	3	0	intron	5

Collection of 18 identified potential off-targets from the six available datasets which appeared to be 5-fold or more downregulated. The first column shows the alignments as found by RNAiSelect which would possibly not have been identified by online available alignment tool. The targeted genes as well as the off-targeted genes predicted by RNAiSelect are listed. The number of regular mismatches and the number of G:U mismatches is given for each off-target. It is listed whether the off-target sequence is present within intron or exon containing sequences of the gene. The fold downregulation compared to the control group (as derived from the available dataset) is presented for each predicted off-target. Functional comparison (using the UniProt Protein knowledgebase; http://www.uniprot.org) did not indicate any functional relation between the targeted gene and these 18 off-targeted genes.

Supplementary Table 4

Targeted gene	Targeted gene description (Uniprot)	Off-targeted gene	Off-target gene description (Uniprot)
CG3941	mitosis; DNA endoreduplication; DNA replication	CG3629	Transcription factor that plays a role in larval and adult appendage development.
CG11184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	CG4128	Ionic channel
CG11184		CG1133	Transcription factor essential for parasegmental subdivision of the embryo.
CG11184		CG12290	G-protein coupled receptor protein signaling pathway
CG8954	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	CG2507	Putative epidermal cell surface receptor
CG2253	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	CG4128	Ionic channel
CG2253		CG2507	Putative epidermal cell surface receptor
CG2253		CG3315	Belongs to the thioredoxin family.
CG2253		CG9656	Transcription factor that is vital to the development of multiple organ systems.
CG2253		CG3234	Forms a heterodimer with period (PER); the complex then translocates into the nucleus. Required for the production of circadian rhythms.
CG2253		CG32130	Apoptosis
CG32743	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	CG13185	Hydrolase
CG32743		CG7978	This is a membrane-bound, calmodulin-insensitive adenylyl cyclase
CG32743		CG12819	nucleolus organization and biogenesis

CG32743		CG4678	Carboxypeptidase
CG32743		CG3359	Unknown
CG1559	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	CG15295	protein binding
CG1559		CG32046	Unknown

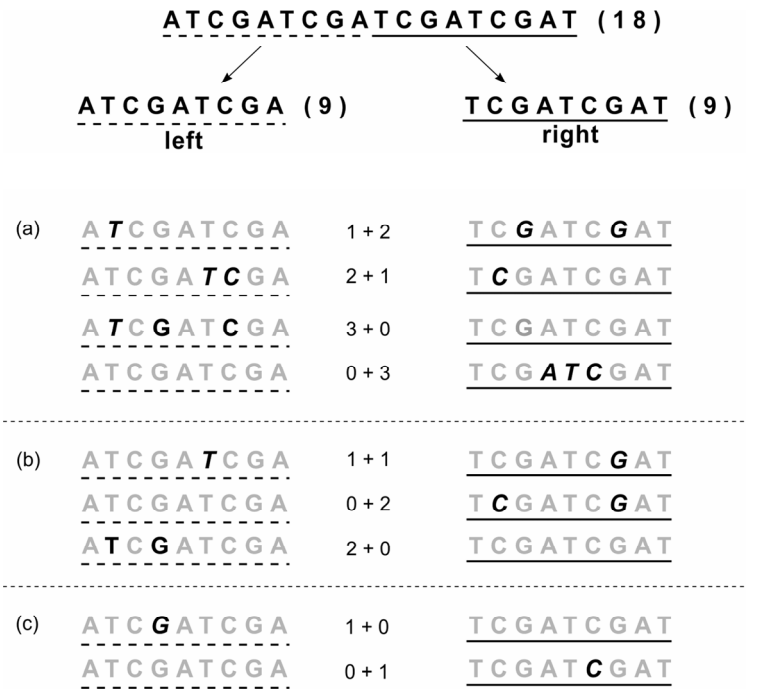
List of functions (as defined by UniProt) of the on-targeted genes from the 6 analyzed dsRNAs and the 18 potential off-target genes listed in Supplementary Table 3.

Supplementary Table 5

Gene	Forward primer	Reverse primer
CG2253	ATGCTAGCCAACGATTCT	CCAGGGCAATCAAATGCA
CG32743	ATGAAGAACGCGATCCAC	GGAGGGCCATGATCATGT
CG8954	ATGGAGGTGACATTCAGC	TGCTTAGTTTGCTGTCGA
CG11184	GTGCCATCTCTATCGGTT	GCTTCCGCTTCTCCTCGT
CG1559	ATGAGCGTGGACACGTACG	TTTGGCGAGCTCGCAGCT
CG3941	GCAGATGTGCAAGCGGGC	TCTCGCACAGGAGACACT

Primers to construct the dsRNAs used in the micro-array experiments.

Supplementary Figure 1



Schematic overview of the distribution of mutations (in bold) along a split 18-nt sequence. (a) The distribution of 3 mismatches (mm) is described by having either 0 mm left and 3 mm right, 1 mm left and 2 mm right, 2 mm left and 1 mm right, or 3 mm left and 0 mm right. (b) The distribution of 2 mismatches is described by having either 0 mm left and 2 mm right, 1 mm left and 1 mm right, or 2 mm left and 0 mm right. (c) The distribution of 1 mismatch is described by having either 0 mm left and 1 mm right, or 1 mm left and 0 mm right.

Chapter 3

RNAi experiments in *D. melanogaster*: solutions to the overlooked problem of off-targets shared by independent dsRNAs

Erwin Seinen, Johannes G.M. Burgerhof, Ritsert C. Jansen, Ody C.M. Sibon

Published in PLoS ONE, October 2010

Abstract

Background: RNAi technology is widely used to downregulate specific gene products. Investigating the phenotype induced by downregulation of gene products provides essential information about the function of the specific gene of interest. When RNAi is applied in *Drosophila melanogaster* or *Caenorhabditis elegans,* often large dsRNAs are used. One of the drawbacks of RNAi technology is that unwanted gene products with sequence similarity to the gene of interest can be down regulated too. To verify the outcome of an RNAi experiment and to avoid these unwanted off-target effects, an additional non-overlapping dsRNA can be used to down-regulate the same gene. However it has never been tested whether this approach is sufficient to reduce the risk of off-targets.
Methodology: We created a novel tool to analyse the occurance of off-target effects in *Drosophila* and we analyzed 99 randomly chosen genes.
Principle findings: Here we show that nearly all genes contain non-overlapping internal sequences that do show overlap in a common off-target gene.
Conclusion: Based on our *in silico* findings, off-target effects should not be ignored and our presented on-line tool enables the identification of two RNA interference constructs, free of overlapping off-targets, from any gene of interest.

Introduction

Genes can be silenced using RNA interference (RNAi). This powerful method is widely used to study biological consequences induced by the down-regulation of selected genes [1-4]. Since its discovery, a great amount of valuable information has been collected using this technology. However, RNAi technology also has some drawbacks such as off-target effects [5-12]. Off-target effects are caused by short stretches of sequence similarity between the RNAi molecule and one or more genes other than the target. Because of high success rates, the fly and worm (*D. melanogaster* and *C. elegans*) model systems generally use large double strand RNAs (dsRNAs) of 300-800 bp. From (large) dsRNAs, numerous siRNAs are generated by the action of DICER and each of these can provoke an RNAi response and exert their gene down-regulating action [13]. Although this results in a favourable synergistic RNAi response towards the target gene, it may in theory also increase the number of off-target possibilities.

A straightforward method to reduce off-target effects, is to use 2 independent and non-overlapping dsRNAs to down-regulate a specific target. Because these dsRNAs are different in sequence composition, their individual off-targets are also assumed to be unique while they both silence the same on-target gene. Consequently, it is reasonable to assume that any shared phenotype which is observed after the independent use of both dsRNAs is an effect of down-regulating the on-target gene (Figure 1A). Although, this line of reasoning is rational, hypothetically it is possible that different non-overlapping siRNAs may actually target different sequences within one identical off-target gene (illustrated in Figure 1B). In such an unfortunate case, a shared off-target effect induced by 2 independent dsRNAs may be misinterpreted as an on-target effect. It has never been investigated what the occurrences are of shared off-target effects when dsRNA are randomly chosen. Here, we present a detailed analysis, based on sequence similarity and a randomized trial which suggest that most genes have independent dsRNA-spanning sequences showing sequence similarity with the same off-target gene. In addition, we present an on-line tool that allows to scan *Drosophila* gene sequences for the occurrence of off-target overlapping regions and to design dsRNAs that have a reduced likelihood to induce identical off-target effects.

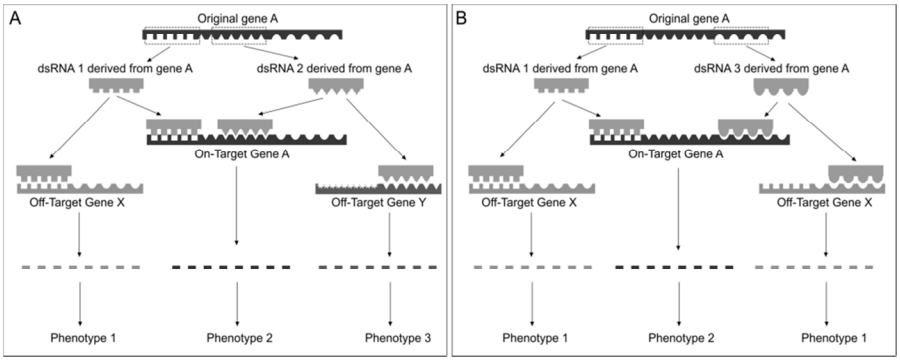


Figure 1 - A: ***Schematic presentation of the event in which identical phenotypes are induced because of shared on-target effects and at the same time different phenotypes are induced because of off-target effects. Phenotype 2 is due to down-regulation of the on-target gene and is induced by dsRNA1 and dsRNA2. Phenotype 1 and 3 are due to down regulation of the off-target gene X and Y respectively and are specific for the individual distinct dsRNAs. In this fortunate event, the individual off-target effects are not identical and are classified as off-target-effects; bona fide conclusions will be drawn from the outcome of this experiment.*** *B*: ***Schematic presentation of the event in which identical phenotypes are induced because of shared on-target effects but at the same***

time an additional identical phenotype is induced by the use of the two independent dsRNAs caused by off-target effects. Phenotype 2 is due to down-regulating the on-target gene and is shared by dsRNA1 and dsRNA3. Phenotype 1 is due to down regulation of a shared off-target gene of the distinct dsRNAs. In this unfortunate event, the off-target effects are identical and will be classified as on-target effects; false conclusions will be drawn from the outcome of this experiment.

Results and Discussion

Statistical analysis on a randomized genome shows that it is likely that 2 distinct 21 nt sequences from the same gene can map closely elsewhere on the genome (see Supplementary file). This hypothetical event (illustrated in Figure 1B) may cause distinct dsRNAs to have common off-targets and that particular combinations of dsRNAs should therefore be avoided. These calculations are based on a non-organized genome containing random sequences, while the *Drosophila* genome is highly functional and far from 'randomized'. To evaluate the risks of our hypothetical event more pragmatically, we used the following approach. First, we picked a dataset of 99 random chosen genes (see supplementary table 1) from the *D. melanogaster* genome. We investigated the occurrence of independent dsRNAs derived from one gene to have shared off-targets. dsRNAs are often derived from cDNA so for our analysis only the cDNA of the 99 genes were considered. Because the complete cDNA can be used to design dsRNAs from, and the dsRNAs are split into siRNAs of approximately 21nt by the RNAi machinery, we first created a list of all possible siRNA sequences that can be obtained from the cDNA sequences of each of these 99 genes. This complete list was subsequently reduced using established scoring rules to exclude 21-bp siRNAs that are most likely non-active (see supplementary table 3). We like to stress that this assumption will only underestimate our findings. Next we calculated the occurrence of all siRNA derived from one cDNA to have a shared off-target with another siRNA derived from the same gene. For this analysis we included pre-mRNA sequences of the complete *D. melanogaster* genome because of the following published results: 1) It has been demonstrated that the RNAi machinery can target pre-mRNAs in *C. elegans* [14]. 2) RNA silencing components have been shown to localize in the nucleus in other organisms (including human) [15-22], further suggesting that pre-mRNAs can be targetted by the RNAi machinery. 3) RNAi constructs can be complementary to miRNAs which are often derived from introns [23] and might act like antagomirs [24]. We therefore analysed the filtered list of

siRNA sequences against both mature and pre-mature RNA sequences to map all possible off-targets with up to 3 mismatches in their sequence alignments with the use of a new tool (see Methods and http://www.RNAiSelect.info/dsrna). By doing so, a list of potential off-targets for each of the individual genes was generated. Next we analyzed whether there was overlap between the potential off-targets of siRNAs derived from the same gene (see Methods). We used the term *cot-group* (Common Off-Target group); a cot-group consists of 2 or more siRNAs, derived from a single gene, that map to the same off-target gene (also illustrated in Table 1; the lines represent members of cot-groups) . The generated siRNA lists of all 99 genes were scanned individually for the presence of cot-groups. The occurrence of cot-groups appeared to be present in all genes, with sometimes excessive high frequencies (Figure 2). As expected, the number of cot-groups are highly correlated with the length of the cDNA of the gene; the larger the sequence, the more cot-groups are formed (Figure 3).

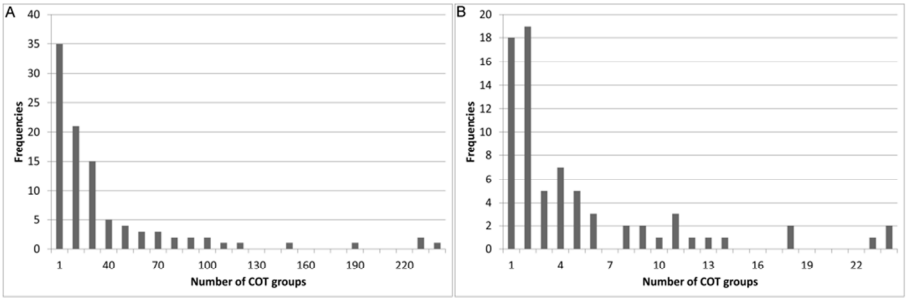


Figure 2 - ***Cot-group frequency distribution for all 99 analyzed genes. A cot-group consists of 2 or more siRNAs, derived from a single gene, that map to the same off-target gene. The x-axis shows the number of cot-groups, the y-axis shows the frequency of genes that contain that number of cot-groups. A: On average, 42 cot-groups were found per gene with a large spread (SD=48.579; N=99) because of the high correlation with gene length (see figure 3). B: Filtering for introns shows that the COT group frequencies are much lower, mostly concentrating around 1-3. Table 1 and Supplementary table 1 illustrate the cot-groups per gene. Table 2 and Supplementary table 2 illustrate the cot-groups per gene after filtering for intron targets.***

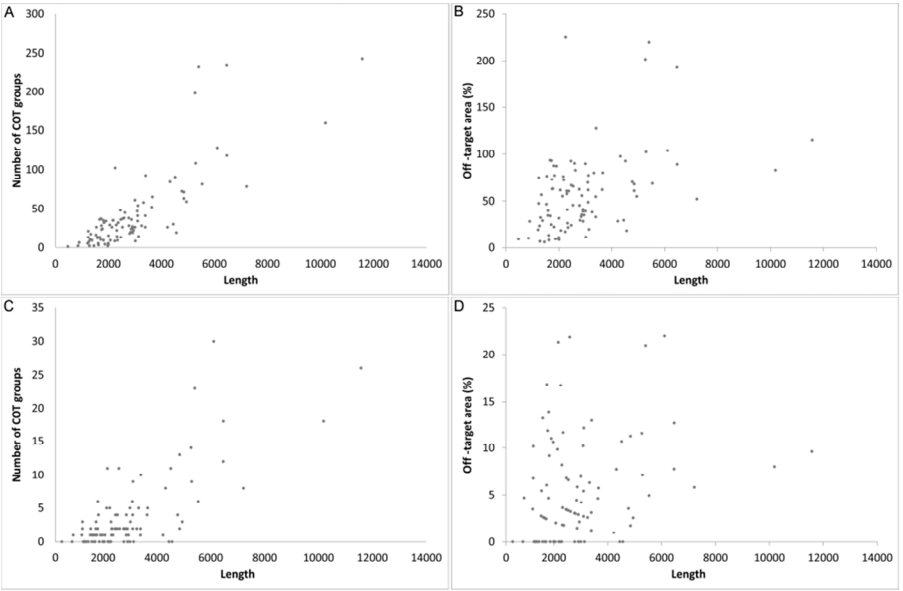


Figure 3 ***- As expected, the length of the gene correlates with the number of cot-groups that can be formed (Spearman's correlation coefficient of 0.44; P < 0.001). This is because the number of potential siRNAs that may originate from one cDNA sequence increases proportionately with the length of the gene. Each potential siRNA adds up on the possibility to form a cot-group. Figure 2A plots the number of cot-groups against the length of the gene. Figure 2B shows the percentage of the sequence within the analyzed genes that map to common off-targets. These values were acquired by multiplying the members of the different cot-groups with the length of a single siRNA (21-bp) within every gene and this is divided by the whole gene length. These values become less reliable with large genes, because there is an increased chance of a single siRNA being part of multiple cot-groups. In that case the area may reach beyond the 100%. Figure 2C and 2D are the results after filtering out intron targets.***

CG2248-RA	1805	30	86%	5	14%	0	0%	0	0%	0	0%	0	0%	35
CG18292-RA	1812	8	89%	1	11%	0	0%	0	0%	0	0%	0	0%	9
CG11372-RA	1832	30	94%	1	3%	1	3%	0	0%	0	0%	0	0%	32
CG5725-RA	1866	4	100%	0	0%	0	0%	0	0%	0	0%	0	0%	4
CG30110-RA	1908	27	79%	6	18%	0	0%	0	0%	0	0%	1	3%	34
CG32669-RA	1977	1	50%	0	0%	0	0%	0	0%	0	0%	1	50%	2
CG1605-RA	1992	4	100%	0	0%	0	0%	0	0%	0	0%	0	0%	4
CG2061-RA	1993	5	100%	0	0%	0	0%	0	0%	0	0%	0	0%	5
CG2720-RA	2000	15	100%	0	0%	0	0%	0	0%	0	0%	0	0%	15
CG11620-RA	2004	4	100%	0	0%	0	0%	0	0%	0	0%	0	0%	4
CG11430-RA	2051	27	93%	2	7%	0	0%	0	0%	0	0%	0	0%	29
CG2522-RA	2074	9	100%	0	0%	0	0%	0	0%	0	0%	0	0%	9
CG11123-RA	2135	23	88%	2	8%	1	4%	0	0%	0	0%	0	0%	26
CG15862-RA	2166	30	86%	2	6%	3	9%	0	0%	0	0%	0	0%	35
CG11450-RA	2185	31	89%	3	9%	1	3%	0	0%	0	0%	0	0%	35
CG1916-RA	2239	7	88%	0	0%	1	13%	0	0%	0	0%	0	0%	8
CG30325-RA	2253	74	73%	22	22%	5	5%	0	0%	0	0%	1	1%	102
CG2674-RA	2294	22	76%	6	21%	1	3%	0	0%	0	0%	0	0%	29
CG11186-RA	2299	26	90%	2	7%	0	0%	1	3%	0	0%	0	0%	29
CG11140-RA	2317	21	95%	1	5%	0	0%	0	0%	0	0%	0	0%	22
CG18455-RA	2325	13	93%	1	7%	0	0%	0	0%	0	0%	0	0%	14
CG5779-RA	2344	7	78%	1	11%	0	0%	0	0%	0	0%	1	11%	9
CG3048-RA	2370	28	88%	4	13%	0	0%	0	0%	0	0%	0	0%	32
CG4822-RA	2453	38	79%	8	17%	2	4%	0	0%	0	0%	0	0%	48
CG12017-RA	2466	32	89%	2	6%	2	6%	0	0%	0	0%	0	0%	36
CG8411-RA	2467	11	85%	2	15%	0	0%	0	0%	0	0%	0	0%	13
CG12178-RA	2531	23	88%	3	12%	0	0%	0	0%	0	0%	0	0%	26
CG8390-RA	2539	32	86%	5	14%	0	0%	0	0%	0	0%	0	0%	37
CG5834-RA	2591	28	74%	3	8%	0	0%	1	3%	1	3%	5	13%	38

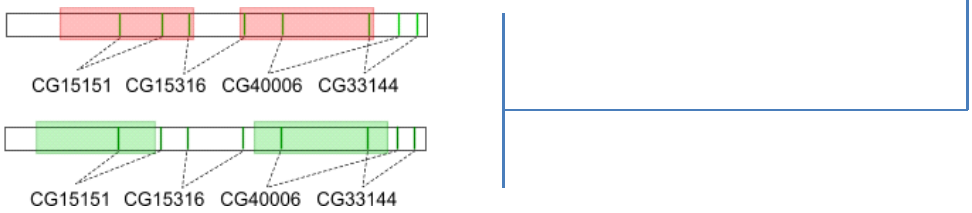


Table 1 - Non-overlapping sequences have a high prevalence of sharing off-targets. *Example of genes for which –based on sequence similarities- overlapping off-targets exist. The number of items per off-target group is given both in numbers and percentages. As an example: the cDNA of gene CG11372-RA contains 30 events of duplicate sequences with a shared off-target and 1 event of triplicate sequences that share the same off-target and 1 event of quadruple sequences that share the same off-target. Green lines represent sites that share an identical off-target with one other site, red lines represent sites that share an identical off-target with 2 other sites, and blue lines with 3 other sites. Purple lines represent sites that share identical off-targets with 5 or more other sites. The complete report from the 99 randomly selected genes are presented in Supplementary table 1. Note that for some genes the lines representing the off-target events are in close proximity and cannot be distinguished as separate lines in this illustrative figure. Overall, there appears to be a tendency for the occurrence of overlapping off-targets at the boundary (UTR's) of the genes, as is evident in for example CG5834 (last gene in the list). The insert shows a more detailed illustration for the analysis of the cDNA of gene CG11620. The green vertical lines represent sites that share an identical off-target with one other site. Sites that share the same off-target are connected with dotted lines and the shared off-target (as CG number) is indicated for each pair. To avoid off-target effects, dsRNA constructs should be chosen in such a way that the dsRNA constructs do not include both members of one pair. The green boxes represent areas which do not include both members of one pair. In order to reduce the likelihood of shared-off target effects, dsRNAs should be designed using sequences from the green regions. In contrast the red areas do include both members of one pair. When 2*

independent dsRNA constructs will be designed from these areas, these dsRNA constructs do share sequence silimalities with the same off-target gene. Our tool provides for all the genes present in the Drosophila genome the green areas.

CG2248-RA	1805	2	100%	0	0%	0	0%	0	0%	0	0%	0	0%	2
CG18292-RA	1812	6	100%	0	0%	0	0%	0	0%	0	0%	0	0%	6
CG11372-RA	1832	4	100%	0	0%	0	0%	0	0%	0	0%	0	0%	4
CG5725-RA	1866	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG30110-RA	1908	0	0%	0	0%	0	0%	0	0%	0	0%	1	100%	1
CG32669-RA	1977	0	0%	0	0%	0	0%	0	0%	0	0%	1	100%	1
CG1605-RA	1992	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG2061-RA	1993	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG2720-RA	2000	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG11620-RA	2004	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG11430-RA	2051	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG2522-RA	2074	1	100%	0	0%	0	0%	0	0%	0	0%	0	0%	1
CG11123-RA	2135	5	100%	0	0%	0	0%	0	0%	0	0%	0	0%	5
CG15862-RA	2166	11	100%	0	0%	0	0%	0	0%	0	0%	0	0%	11
CG11450-RA	2185	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG1916-RA	2239	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG30325-RA	2253	4	80%	0	0%	0	0%	0	0%	0	0%	1	20%	5
CG2674-RA	2294	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG11186-RA	2299	2	67%	0	0%	0	0%	1	33%	0	0%	0	0%	3
CG11140-RA	2317	1	100%	0	0%	0	0%	0	0%	0	0%	0	0%	1
CG18455-RA	2325	2	100%	0	0%	0	0%	0	0%	0	0%	0	0%	2
CG5779-RA	2344	0	0%	1	50%	0	0%	0	0%	0	0%	1	50%	2
CG3048-RA	2370	1	100%	0	0%	0	0%	0	0%	0	0%	0	0%	1
CG4822-RA	2453	2	100%	0	0%	0	0%	0	0%	0	0%	0	0%	2
CG12017-RA	2466	4	100%	0	0%	0	0%	0	0%	0	0%	0	0%	4
CG8411-RA	2467	2	100%	0	0%	0	0%	0	0%	0	0%	0	0%	2
CG12178-RA	2531	2	100%	0	0%	0	0%	0	0%	0	0%	0	0%	2
CG8390-RA	2539	4	100%	0	0%	0	0%	0	0%	0	0%	0	0%	4
CG5834-RA	2591	9	82%	0	0%	1	9%	1	9%	0	0%	0	0%	11

Table 2 – Non-overlapping sequences have a significant prevalence of sharing exon targeting off-targets. *All 99 cDNA sequences were re-analyzed, now ignoring any predicted off-targets that occur within intron sequences. Although this lowers the predicted off-targets, it still shows a significant occurrence of overlapping off-targets. Our analysis shows that after applying this filter, 74% of the genes show off-targets that occur more than once within the same derived cDNA sequence.*

We then looked at the general profile of the cot-groups for each gene separately and tried to deduce the required number of dsRNAs to strongly reduce the event of common off-targets. If for example the number of members within the cot-groups does not exceed 2, than this implies that at most 2 siRNAs within the same gene map to a common off-target. For that particular situation, the use of 3 or more non-overlapping dsRNA will always generate bona fide data as there is no possibility for all 3 of them to share a common off-target (bases on sequence similarity). Unfortunately, most genes have cot-groups with at least 3 members (Figure 4, also depicted by red lines in Table 1 and in Supplementary table 1) or even 4 members (Figure 4, also depicted by blue lines in Table 1 and in Supplementary table 1). This finding demonstrates that just using multiple non-overlapping dsRNAs is not sufficient to exclude off-target events (see also insert Table 1), even if the number of independent dsRNAs is 3 or more. We therefore developed a bioinformatics approach to design dsRNAs that avoids all predicted off-targets. Our freely available website presents such a tool at http://www.RNAiSelect.info/dsrna. This web based tool accepts a gene name as input and presents a number of choices each containing a combination of 2 unique dsRNAs that lack overlapping –based on sequence similarity- off-targets (Supplementary Figure 1).

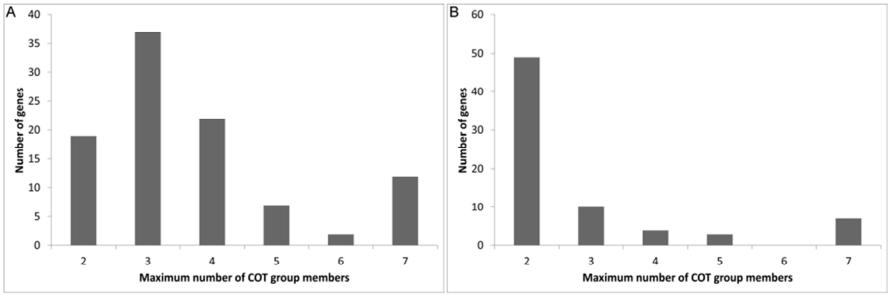


Figure 4 - ***Distribution of the maximum cot-group size for the 99 analyzed genes. For each cot-group with the indicated number of members on the x-axis, the number of genes where counted that have cot-groups with a corresponding maximum cot-group size. This shows for example that in figure 4A, there are 19 genes that have cot-groups with no more than 2 members. Most genes (38) in our analysis appear to have at least one cot-group present with up to 3 members. Due to the gene length with cot-group correlation, some large genes in our analysis also show very large cot-groups which account for the unexpected large 7+ count as plotted in the last bar (see also Table 1 and Supplementary Table 1). Figure 4B presents the results after filtering out intron targets, showing that most genes have at least one cot-group present with 2 members.***

Next, we repeated the above analysis, but now only considering off-targets targeting mature RNA sequences, because these are maybe be more active in RNAi [25]. Filtering out the intron off-targets, causes much less off-targets to be found in general per cDNA (Table 2). Overall, both the sizes and occurrences of the COT groups are smaller (Figure 2B, Figure 4B). Nevertheless, there is still a significant number of overlapping potential off-targets to be expected in >74% of the genes. In 24% of the analysed genes there is at least 1 COT groups present of size 3 (Figure 4B), meaning that there are at least 3 areas within the cDNA that target the same off-target. Therefore, even if only mature RNA sequences are considered and 2 randomly chosen non-overlapping dsRNA's are used, the experimental outcome can be obscured by off-target effects. This further underscores the utility of our tool.

Although not exclusively, we observed a strong tendency for overlapping off-targets to occur at the end of genes (see graphical illustrations of the common off-targets in Table 1 and in Supplementary Table 1), corresponding to the untranslated regions (UTR). The UTR sequences are less unique in the genome as compared to the coding region and therefore preferably should be avoided when dsRNA constructs are designed. Our tool includes an option to avoid UTR sequences to minimize off-targets when designing dsRNAs of interest.

Conclusion

Our analysis demonstrated that most genes in the *D. melanogaster* genome contain 2 or more (distinct) sequences that show sequence similarity (containing 3 or less mismatches) to the same off-target gene. The potential consequence of these overlapping occurrences is that 2 dsRNAs which are generated to down-regulate a specific target gene, may possess a common off-target gene as well. In case these 2 distinct dsRNAs are used, their common phenotype induced by down-regulation of their shared off-target gene may lead to misinterpretation of the experiment. We present a method to identify 2 distinct dsRNAs from a gene of choice that do not show any off-target overlap, -based on sequence similarity- by performing a thorough off-target overlap analysis. This tool is freely available at http://www.rnaiselect.info/dsrna and may be used the *Drosophila* community where dsRNAs are generally used for gene down-regulation.

Methods

Genomic data (build 45-43b for *Drosophila*) were downloaded from the Ensembl website (www.ensembl.org). The data from Ensemble and its derived seed tables were processed, stored and indexed in a MySQL database, version 5.0, running on top of Ubuntu 6.06. The on-line available RNAiSelect program (http://www.RNAiSelect.info/) was written in C#.NET and performs a comprehensive sequence alignment against the input genome for up to 3 mismatches.

The RNAiSelect algorithm was specifically designed for finding relationships between short nucleotide sequences. It has a high performance and usability for short-sequence studies, including siRNA (off-)targets. The complete source code and documentation for a standalone version of this algorithm may be downloaded from http://rnaiselect.sourceforge.net/. The algorithm is based on the following assumption:

"An example sequence TTTTAATTTGGGCCCGGG consists of 18 nucleotides and may be split into two 9-nt child sequences; TTTTAATTT and GGGCCCGGG. By plain observation, we know that the sequence GGGCCCGGG is exactly 9-nt separated from TTTTAATTT in the original sequence."

For the RNAiSelect algorithm to work, we first wrote a program that generates a seed table which holds the exact genomic location(s) for every possible 9-nt sequence (4^9, or 262.144 sequences). Generating such an index is a general strategy used by many algorithms to rapidly look-up any sequence of fixed length for its positions in the genome. The used algorithm however uses a novel method to calculate the positional relationship between indexed seeds, instead of performing string-to-string comparisons for every nucleotide after a hit has been found. In other words, by searching 9-nt subsequences of the whole query sequence for consecutive matches of locations, it will find hits larger than 9nt without performing actual DNA comparisons. This following example, in layman code, shows how to find an 18-nt sequence in the genome by first splitting the sequence into its two 9-nt subsequences and comparing these sequences with the available index table with a word size of 9.

1	SPLIT QUERY SEQUENCE(18 nt) INTO *dnacode_left(9 nt)* AND *dnacode_right(9 nt)*
2	EXTRACT LOCATIONS FROM *index_table* FOR dnacode_left AND STORE IN *seedtable_left*
3	EXTRACT LOCATIONS FROM *index_table* FOR dnacode_ *right* AND STORE IN *seedtable_* right
4	SELECT ALL HITS WHERE (LOCATIONS *seedtable_left + 9) EQUALS (LOCATIONS seedtable_right)*

This example merely demonstrates how to find an exact 18-nt hit not allowing any mismatches. However, mismatches may be added by expanding the seed searches with variations so that all possible combinations will be found. We thus included variations of the 9-nt sub-sequences and then compared the distance relationship between the original locations of the seed hits, which has to be exactly 9. This may considerably increase the number of seed searches, but because these are relatively cheap in terms of processing time, the overall performance is very high while it guarantees that every possible alignment is evaluated.

The cDNA sequences from each 99 gene (Supplementary Table 1) were first analyzed for potentially active sequences as might be produced by endogenous DICER. A scoring schema was used (Supplementary Table 2) during this analysis to estimate and extract the most potential sequences. Each extracted sequence was analyzed for potential off-targets. The combined output from all these potential off-targets was cross-referenced with each other to map areas on the original cDNA sequence that are predicted to have overlapping off-targets. At the same time, regions can be identified that lack these areas and dsRNA sequences can be extracted that are completely devoid of off-targets. The results are presented and 2 or more dsRNA are indicated that originate from the same gene and that are predicted to lack overlapping off-targets. From the indicated areas, dsRNA can be designed. An identical analysis can be done for every Drosophila gene of interest through a web-interface at http://www.rnaiselect.info/dsrna which presents the output in a user friendly interface.

References

1. Moazed D. Small RNAs in transcriptional gene silencing and genome defence, Nature 2009;457:413-420.
2. Castanotto D, Rossi JJ. The promises and pitfalls of RNA-interference-based therapeutics, Nature 2009;457:426-433.
3. Dietzl G, Chen D, Schnorrer F et al. A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila, Nature 2007;448:151-156.
4. Morris K, Simon WLC, Jacobsen S et al. Small Interfering RNA-Induced Transcriptional Gene Silencing in Human Cells, Science 2004;305:1289-1292.
5. Qiu S, Adema C, Lane T. A computational study of off-target effects of RNA interference, Nucleic Acids Res 2005;33:1834-1847.
6. Kulkarni M, Booker M, Silver S et al. Evidence of off-target effects associated with long dsRNAs in Drosophila melanogaster cell-based assays, Nat Methods 2006;3:833-838.
7. Moffat J, Reiling JH, Sabatini DM. Off-target effects associated with long dsRNAs in Drosophila RNAi screens, Trends in pharmacological sciences 2007;28:149-151.
8. Fedorov Y, Anderson EM, Birmingham A et al. Off-target effects by siRNA can induce toxic phenotype, RNA 2006;12:1188-1196.
9. Ma Y, Creanga A, Lum L et al. Prevalence of off-target effects in Drosophila RNA interference screens, Nature 2006;443:359-363.
10. Doench JG, Petersen CP, Sharp PA. siRNAs can function as miRNAs, Genes Dev 2003;17:438-442.
11. Saxena S, Jonsson Z, Dutta A. Small RNAs with imperfect match to endogenous mRNA repress translation. Implications for off-target activity of small inhibitory RNA in mammalian cells, The Journal of biological chemistry 2003;278:44312-44319.
12. Jackson A, Burchard J, Schelter J et al. Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity, RNA 2006;12:1179-1187.
13. Hannon GJ. RNA interference, Nature 2002;418:244-251.
14. Bosher J, Dufourcq P, Sookhareea S et al. RNA Interference Can Target Pre-mRNA: Consequences for Gene Expression in a Caenorhabditis elegans Operon, Genetics 1999;153:1245-1256.
15. Pal-Bhadra M, Bhadra U, Birchler JA. RNAi Related Mechanisms Affect Both Transcriptional and Posttranscriptional Transgene Silencing in Drosophila, Molecular cell 2002;9:315-327.
16. Verdel A, Jia S, Gerber S et al. RNAi-Mediated Targeting of Heterochromatin by the RITS Complex, Science 2004;303:672-676.

17. Langlois M-A, Boniface C, Wang G et al. Cytoplasmic and Nuclear Retained DMPK mRNAs Are Targets for RNA Interference in Myotonic Dystrophy Cells, J Biol Chem 2005;280:16949-16954.
18. Matzke MA, Birchler JA. RNAi-mediated pathways in the nucleus, Nat Rev Genet 2005;6:24-35.
19. Robb GB, Brown KM, Khurana J et al. Specific and potent RNAi in the nucleus of human cells, Nat Struct Mol Biol 2005;12:133-137.
20. Weinberg MS, Barichievy S, Schaffer L et al. An RNA targeted to the HIV-1 LTR promoter modulates indiscriminate off-target gene activation, Nucleic Acids Res 2007;35:7303-7312.
21. Lin S-L, Kim H, Ying S-Y. Intron-mediated RNA interference and microRNA (miRNA), Frontiers in bioscience : a journal and virtual library 2008;13:2216-2230.
22. Politz JC, Hogan EM, Pederson T. MicroRNAs with a nucleolar location, RNA 2009;15:1705-1715.
23. Lin SL, Miller JD, Ying SY. Intronic MicroRNA (miRNA), J Biomed Biotechnol 2006;2006:26818.
24. Krutzfeldt J, Rajewsky N, Braich R et al. Silencing of microRNAs in vivo with 'antagomirs', Nature 2005;438:685-689.
25. Elbashir SM, Lendeckel W, Tuschl T. RNA interference is mediated by 21- and 22-nucleotide RNAs, Genes & development. 2001;15:188-200.
26. Adams J. Gene duplication and the birthday problem, Nature 1982;296:176-176.
27. Holen T, Moe S, Sorbo J et al. Tolerated wobble mutations in siRNAs decrease specificity, but can enhance activity in vivo, Nucleic Acids Res 2005;33:4704-4710.
28. Wakiyama M, Matsumoto T, Yokoyama S. Drosophila U6 promoter-driven short hairpin RNAs effectively induce RNA interference in Schneider 2 cells, Biochem Biophys Res Commun 2005;331:1163-1170.
29. Aleman L, Doench J, Sharp P. Comparison of siRNA-induced off-target RNA and protein effects, RNA 2007;13:385-395.
30. Haley B, Zamore PD. Kinetic analysis of the RNAi enzyme complex, Nature Structural & Molecular Biology 2004;11:599-606.

Supporting Information Chapter 3

To investigate the probability of shared off-targets, we first addressed the question whether a random sequence will likely match or nearly match to any other sequence present in the *D. melanogaster* genome. To estimate this, a formula previously designed by others for the same purpose [26], was used. With this formula we calculated what the occurrence is of any random sequence of 21 nt to find one other (k=2) identical sequence within the genome. For this calculation, the 21 nt sequence and the complete genome are for the sake of simplicity considered to consist of random sequences. In this formula, *n* is the number of possible different sequences of a specific sequence length, equal to 4^{21} for a sequence of 21 nt long (assuming 4 possible different nucleotides; A, T, C or G). 21 nt was chosen because siRNAs of this size may be generated when large dsRNAs are used in RNAi experiments for the *D. melanogaster* genome. The number of copies to be present is represented by k (=2).

(1) $$\boldsymbol{\lambda = n \frac{e^{-\frac{r}{n}}}{k!} \left(\frac{r}{n}\right)^{k}}$$

The size of the Drosophila genome is 168.7 Mb (r), therefore it can be expected that, within a randomized genome of this size, there are 3,235 occurrences (λ) of random 21 nt sequence that exactly match to another sequence within that same genome. However, in this estimation mismatches are neglected while sequences containing mismatches may also contribute to a measurable RNAi response [27-30]. When mismatches are taken into account, this considerably increases the number of different possibilities for an siRNA to find successful sequence hybridization partners with other mRNAs. The number of sequences that are possible with length *n* and up to *k* mismatches is calculated with the standard formula:

(2) $$\frac{\boldsymbol{n!}}{\boldsymbol{k!\cdot(n-k)!}}$$

Note that in this standard formula the symbols *n* and *k* are also used, however they are different from *n* and *k* presented in formula 1. We used 3 mismatches in this and following calculations as this has been known to elicit a successful RNAi response [6]. Using *k=3* results in 1,330 different combinations for a 21-bp sequence. Combining this with formula (1), *n* is divided by 1,330 and λ is recalculated to be 4,089,178 occurrences instead of the previous calculated 3,235 for exact matches. Each occurrence may span up to 42-nt (2x the size of a single occurring siRNA; 21-nt), and therefore all events together (42 x 4,089,178) may cover the complete genome.

Supplementary Figure 1

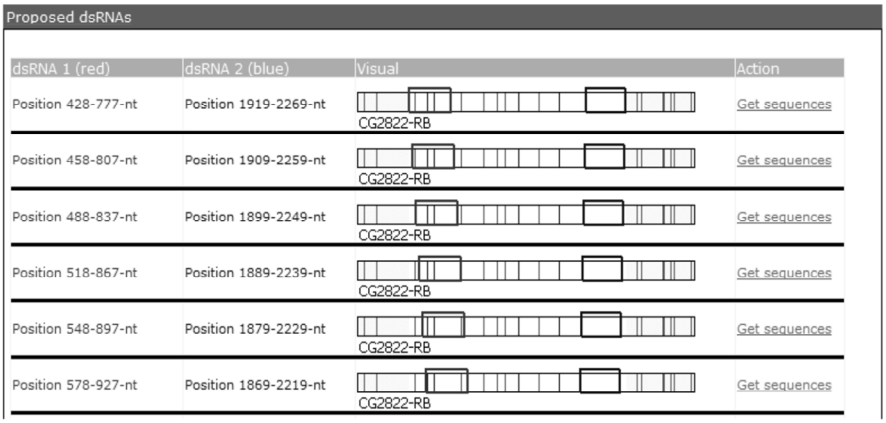


Illustration of the on-line website (www.rnaiselect.info/dsrna), enabling the identification of dsRNA constructs that do not have shared off-targets. The table lists different dsRNAs combinations (red and blue boxes) that do not have predicted overlapping off-targets (off-targets are indicated by vertical lines) on top of a schematic view of the complete cDNA sequence. By using these dsRNA sequences, the chances that the experimental outcome is influenced by shared off-target effects will most likely be reduced.

Supplementary Table 1

Description	Score
30%-52% GC Content	1 point
3 or more A/Us at positions 15-19	1 point per A/U
Tm>20°C	1 point
A at position 19	1 point
A at position 3	1 point
U at position 10	1 point
G/C at position 19	-1 point
G at position 13	-1 point
>4 sequential nucleotide repeat	-9 points
> 4 diplet repeat	-9 points

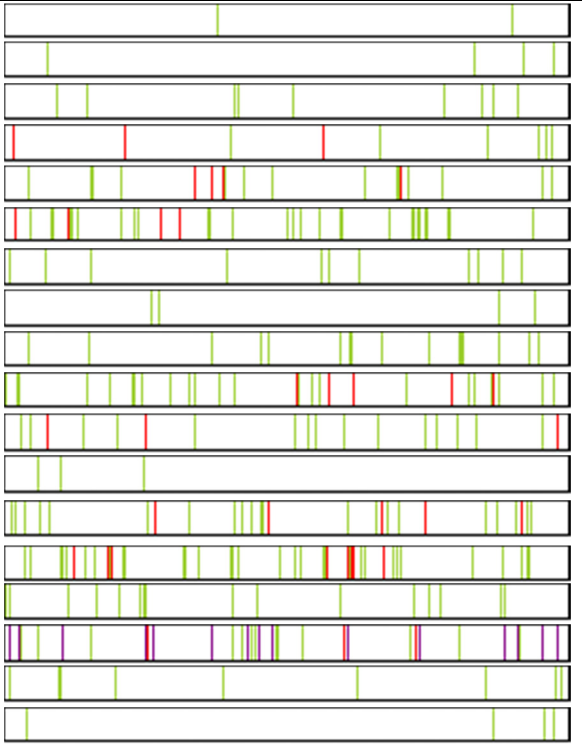
Scoring scheme used to define most potent siRNAs, based on a summary from several publications. Only sequences scoring at least 6 points were considered.

References:

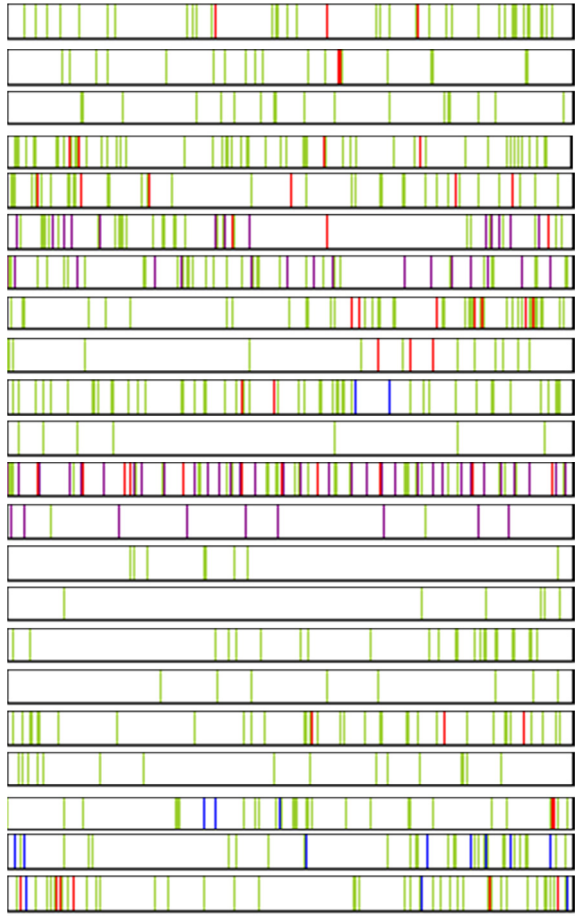
- http://www.protocol-online.org/prot/Protocols/Rules-of-siRNA-design-for-RNA-interference--RNAi--3210.html
- Jagla, B., Aulner, N., Kelly, P.D., Song, D., Volchuk, A., Zatorski, A., Shum, D., Mayer, T., De Angelis, D.A., Ouerfelli, O. et al. (2005) Sequence characteristics of functional siRNAs. RNA, 11, 864-872.
- Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W. and Khvorova, A. (2004) Rational siRNA design for RNA interference. Nature biotechnology., 22, 326-330.
- Elbashir, S.M., Martinez, J., Patkaniowska, A., Lendeckel, W. and Tuschl, T. (2001) Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate. EMBO J, 20, 6877-6888.

Supplementary Table 2

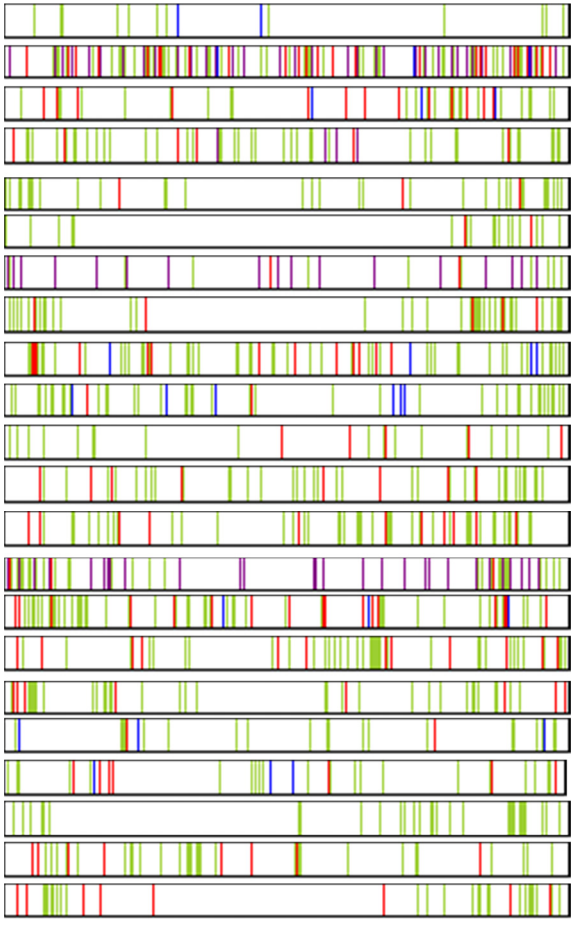
		# of overlapping off-targets												
Gene	**Length (nt)**	**2**	**%**	**3**	**%**	**4**		**5**		**6**		**>6**		**Total**
CG11455-RA	471	1	100%	0	0%	0	0%	0	0%	0	0%	0	0%	1
CG11454-RA	852	2	100%	0	0%	0	0%	0	0%	0	0%	0	0%	2
CG11488-RA	898	6	100%	0	0%	0	0%	0	0%	0	0%	0	0%	6
CG31975-RA	1215	4	80%	1	20%	0	0%	0	0%	0	0%	0	0%	5
CG17691-RA	1238	11	85%	2	15%	0	0%	0	0%	0	0%	0	0%	13
CG10465-RA	1241	19	90%	2	10%	0	0%	0	0%	0	0%	0	0%	21
CG11617-RA	1267	7	100%	0	0%	0	0%	0	0%	0	0%	0	0%	7
CG1712-RA	1284	2	100%	0	0%	0	0%	0	0%	0	0%	0	0%	2
CG8245-RA	1306	10	100%	0	0%	0	0%	0	0%	0	0%	0	0%	10
CG11374-RA	1338	15	88%	2	12%	0	0%	0	0%	0	0%	0	0%	17
CG3436-RA	1381	8	89%	1	11%	0	0%	0	0%	0	0%	0	0%	9
CG3388-RA	1452	2	100%	0	0%	0	0%	0	0%	0	0%	0	0%	2
CG5680-RB	1530	14	88%	2	13%	0	0%	0	0%	0	0%	0	0%	16
CG17684-RA	1546	22	85%	4	15%	0	0%	0	0%	0	0%	0	0%	26
CG11377-RA	1585	13	100%	0	0%	0	0%	0	0%	0	0%	0	0%	13
CG11125-RA	1591	10	83%	1	8%	0	0%	0	0%	0	0%	1	8%	12
CG11127-RA	1617	5	100%	0	0%	0	0%	0	0%	0	0%	0	0%	5
CG2863-RA	1625	3	100%	0	0%	0	0%	0	0%	0	0%	0	0%	3



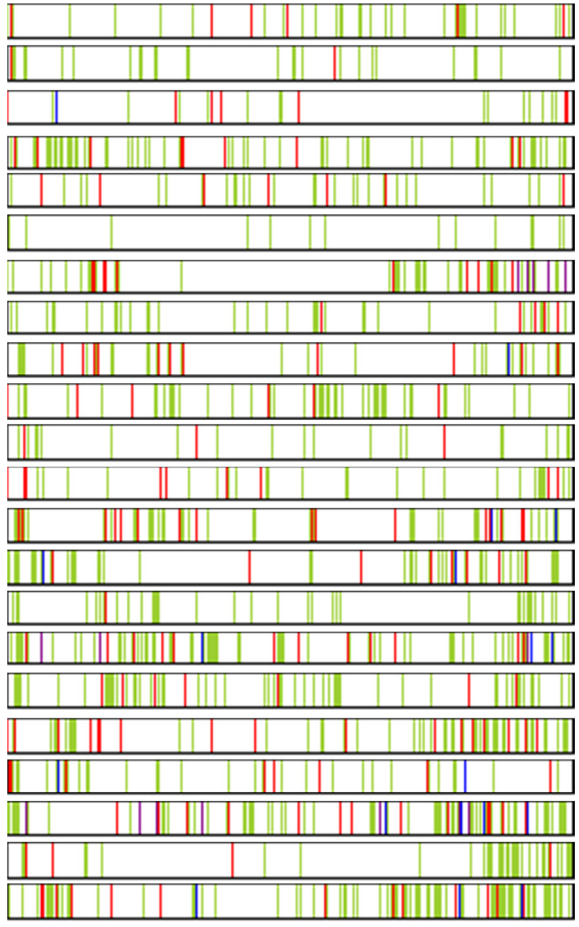
CG1903-RA	1666	34	94%	2	6%	0	0%	0	0%	0	0%	0	0%	36
CG2657-RA	1674	12	92%	1	8%	0	0%	0	0%	0	0%	0	0%	13
CG40449-RA	1707	10	100%	0	0%	0	0%	0	0%	0	0%	0	0%	10
CG3709-RA	1719	35	95%	2	5%	0	0%	0	0%	0	0%	0	0%	37
CG31974-RA	1736	24	86%	4	14%	0	0%	0	0%	0	0%	0	0%	28
CG14489-RA	1744	21	91%	1	4%	0	0%	0	0%	0	0%	1	4%	23
CG11023-RA	1773	23	96%	0	0%	0	0%	0	0%	0	0%	1	4%	24
CG2248-RA	1805	30	86%	5	14%	0	0%	0	0%	0	0%	0	0%	35
CG18292-RA	1812	8	89%	1	11%	0	0%	0	0%	0	0%	0	0%	9
CG11372-RA	1832	30	94%	1	3%	1	3%	0	0%	0	0%	0	0%	32
CG5725-RA	1866	4	100%	0	0%	0	0%	0	0%	0	0%	0	0%	4
CG30110-RA	1908	27	79%	6	18%	0	0%	0	0%	0	0%	1	3%	34
CG32669-RA	1977	1	50%	0	0%	0	0%	0	0%	0	0%	1	50%	2
CG1605-RA	1992	4	100%	0	0%	0	0%	0	0%	0	0%	0	0%	4
CG2061-RA	1993	5	100%	0	0%	0	0%	0	0%	0	0%	0	0%	5
CG2720-RA	2000	15	100%	0	0%	0	0%	0	0%	0	0%	0	0%	15
CG11620-RA	2004	4	100%	0	0%	0	0%	0	0%	0	0%	0	0%	4
CG11430-RA	2051	27	93%	2	7%	0	0%	0	0%	0	0%	0	0%	29
CG2522-RA	2074	9	100%	0	0%	0	0%	0	0%	0	0%	0	0%	9
CG11123-RA	2135	23	88%	2	8%	1	4%	0	0%	0	0%	0	0%	26
CG15862-RA	2166	30	86%	2	6%	3	9%	0	0%	0	0%	0	0%	35
CG11450-RA	2185	31	89%	3	9%	1	3%	0	0%	0	0%	0	0%	35



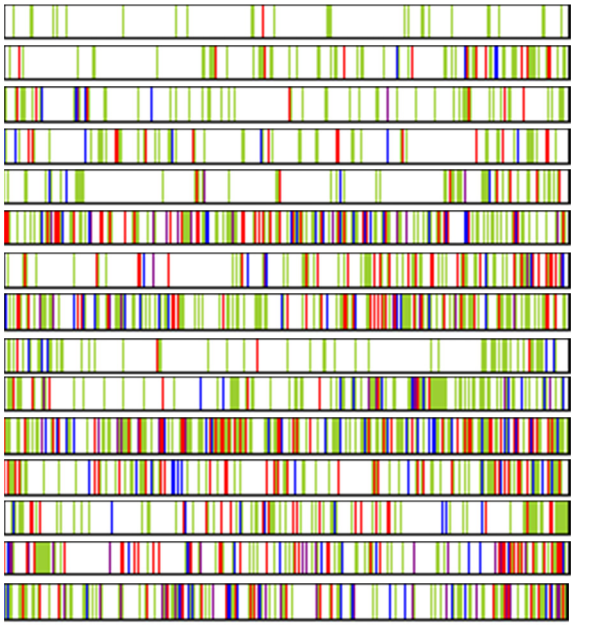
CG1916-RA	2239	7	88%	0	0%	1	13%	0	0%	0	0%	0	0%	8
CG30325-RA	2253	74	73%	22	22%	5	5%	0	0%	0	0%	1	1%	102
CG2674-RA	2294	22	76%	6	21%	1	3%	0	0%	0	0%	0	0%	29
CG11186-RA	2299	26	90%	2	7%	0	0%	1	3%	0	0%	0	0%	29
CG11140-RA	2317	21	95%	1	5%	0	0%	0	0%	0	0%	0	0%	22
CG18455-RA	2325	13	93%	1	7%	0	0%	0	0%	0	0%	0	0%	14
CG5779-RA	2344	7	78%	1	11%	0	0%	0	0%	0	0%	1	11%	9
CG3048-RA	2370	28	88%	4	13%	0	0%	0	0%	0	0%	0	0%	32
CG4822-RA	2453	38	79%	8	17%	2	4%	0	0%	0	0%	0	0%	48
CG12017-RA	2466	32	89%	2	6%	2	6%	0	0%	0	0%	0	0%	36
CG8411-RA	2467	11	85%	2	15%	0	0%	0	0%	0	0%	0	0%	13
CG12178-RA	2531	23	88%	3	12%	0	0%	0	0%	0	0%	0	0%	26
CG8390-RA	2539	32	86%	5	14%	0	0%	0	0%	0	0%	0	0%	37
CG5834-RA	2591	28	74%	3	8%	0	0%	1	3%	1	3%	5	13%	38
CG11665-RA	2621	34	76%	9	20%	2	4%	0	0%	0	0%	0	0%	45
CG3164-RA	2768	32	84%	6	16%	0	0%	0	0%	0	0%	0	0%	38
CG5748-RA	2777	24	86%	4	14%	0	0%	0	0%	0	0%	0	0%	28
CG1616-RA	2801	16	89%	1	6%	1	6%	0	0%	0	0%	0	0%	18
CG1464-RA	2844	16	80%	3	15%	1	5%	0	0%	0	0%	0	0%	20
CG11486-RA	2862	24	100%	0	0%	0	0%	0	0%	0	0%	0	0%	24
CG5036-RA	2869	32	89%	4	11%	0	0%	0	0%	0	0%	0	0%	36
CG4889-RA	2907	21	84%	4	16%	0	0%	0	0%	0	0%	0	0%	25



CG11490-RA	2909	24	89%	3	11%	0	0%	0	0%	0	0%	0	0%	27
CG2144-RA	2914	18	95%	1	5%	0	0%	0	0%	0	0%	0	0%	19
CG32465-RB	2948	18	82%	3	14%	1	5%	0	0%	0	0%	0	0%	22
CG10283-RA	3001	55	90%	6	10%	0	0%	0	0%	0	0%	0	0%	61
CG4648-RA	3003	22	85%	4	15%	0	0%	0	0%	0	0%	0	0%	26
CG2331-RA	3018	8	100%	0	0%	0	0%	0	0%	0	0%	0	0%	8
CG4637-RA	3089	41	87%	5	11%	0	0%	1	2%	0	0%	0	0%	47
CG11166-RA	3103	30	88%	4	12%	0	0%	0	0%	0	0%	0	0%	34
CG11121-RA	3104	33	80%	7	17%	1	2%	0	0%	0	0%	0	0%	41
CG15207-RA	3114	48	89%	6	11%	0	0%	0	0%	0	0%	0	0%	54
CG11579-RA	3136	13	93%	1	7%	0	0%	0	0%	0	0%	0	0%	14
CG9211-RA	3252	25	89%	3	11%	0	0%	0	0%	0	0%	0	0%	28
CG4698-RA	3323	49	84%	8	14%	1	2%	0	0%	0	0%	0	0%	58
CG11066-RA	3391	36	88%	4	10%	1	2%	0	0%	0	0%	0	0%	41
CG18492-RA	3392	25	96%	1	4%	0	0%	0	0%	0	0%	0	0%	26
CG8426-RA	3405	76	83%	12	13%	2	2%	2	2%	0	0%	0	0%	92
CG3836-RA	3637	49	94%	3	6%	0	0%	0	0%	0	0%	0	0%	52
CG3938-RA	3658	56	86%	9	14%	0	0%	0	0%	0	0%	0	0%	65
CG5076-RA	4228	22	85%	3	12%	1	4%	0	0%	0	0%	0	0%	26
CG31973-RA	4324	62	74%	14	17%	5	6%	3	4%	0	0%	0	0%	84
CG2186-RA	4449	28	93%	2	7%	0	0%	0	0%	0	0%	0	0%	30
CG1624-RA	4516	75	83%	11	12%	4	4%	0	0%	0	0%	0	0%	90



CG10079-RA	4563	18	95%	1	5%	0	0%	0	0%	0	0%	0	0%	19
CG1708-RA	4772	59	82%	11	15%	2	3%	0	0%	0	0%	0	0%	72
CG42311-RA	4846	53	84%	7	11%	2	3%	1	2%	0	0%	0	0%	63
CG5098-RA	4846	60	85%	8	11%	3	4%	0	0%	0	0%	0	0%	71
CG4952-RA	4943	52	88%	4	7%	2	3%	1	2%	0	0%	0	0%	59
CG40351-RA	5268	127	64%	46	23%	16	8%	5	3%	3	2%	1	1%	198
CG5753-RA	5293	78	72%	21	19%	6	6%	3	3%	0	0%	0	0%	108
CG2671-RA	5407	165	71%	43	19%	17	7%	5	2%	1	0%	1	0%	232
CG2411-RA	5535	66	81%	12	15%	3	4%	0	0%	0	0%	0	0%	81
CG1725-RA	6109	101	79%	12	9%	12	9%	2	2%	1	1%	0	0%	128
CG11376-RA	6463	152	65%	50	21%	24	10%	5	2%	2	1%	1	0%	234
CG2146-RA	6470	88	75%	22	19%	8	7%	0	0%	0	0%	0	0%	118
CG6383-RA	7216	59	76%	15	19%	4	5%	0	0%	0	0%	0	0%	78
CG3936-RA	10189	104	65%	37	23%	10	6%	6	4%	2	1%	0	0%	159
CG9995-RA	11579	164	67%	40	16%	22	9%	10	4%	1	0%	6	2%	243

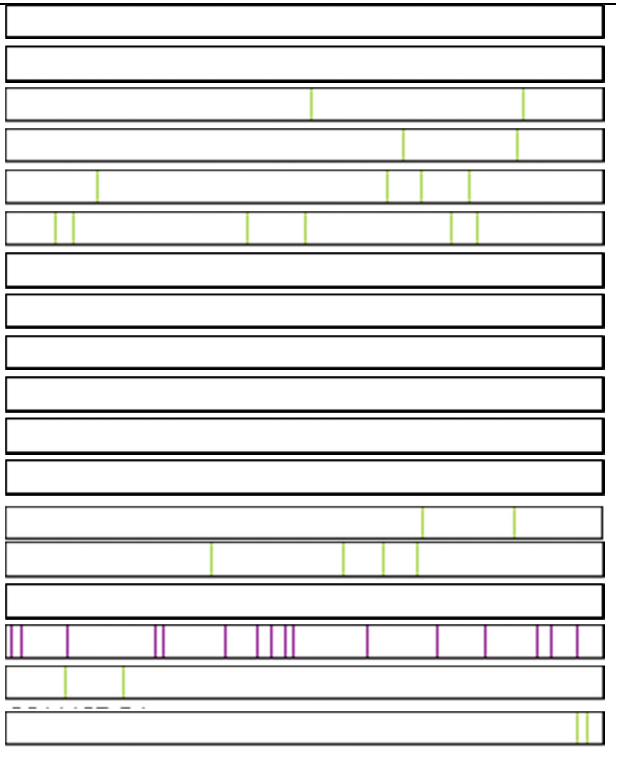


Detailed report of all 99 analyzed genes showing the overlap between sequence similarities based off-targets. The number of items per off-target group is given both in numbers and percentages. As an example: the cDNA of gene CG11372-RA contains 30 events of duplicate sequences with a shared off-target and 1 event of triplicate sequences that share the same off-target and 1 event of quadruple sequences that share the same of target. Green lines represent sites that share an identical off-target with one other site, red lines represent sites that share an identical off-target with 2 other sites, and blue lines with 3 other sites. Purple lines represent sites that share identical off-

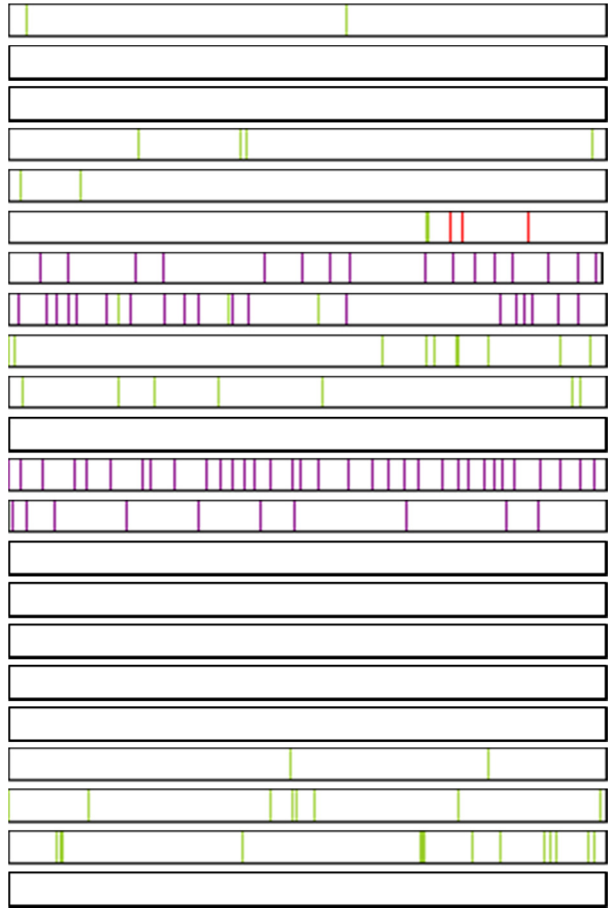
targets with 5 or more other sites. Note that for some genes the lines representing the off-target events are in close proximity and cannot be distinguished as separate lines in this illustrative figure.

Supplementary Table 3

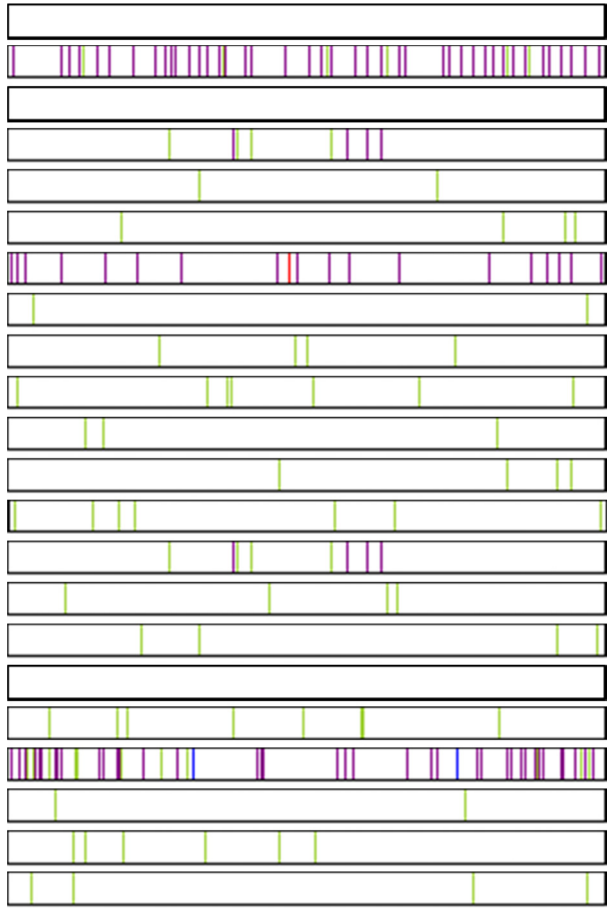
		# of overlapping off-targets												
Gene	**Total length**	**2**	**%**	**3**	**%**	**4**		**5**		**6**		**>6**		**Total**
CG11455-RA	471	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG11454-RA	852	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG11488-RA	898	1	100%	0	0%	0	0%	0	0%	0	0%	0	0%	1
CG31975-RA	1215	1	100%	0	0%	0	0%	0	0%	0	0%	0	0%	1
CG17691-RA	1238	2	100%	0	0%	0	0%	0	0%	0	0%	0	0%	2
CG10465-RA	1241	3	100%	0	0%	0	0%	0	0%	0	0%	0	0%	3
CG11617-RA	1267	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG1712-RA	1284	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG8245-RA	1306	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG11374-RA	1338	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG3436-RA	1381	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG3388-RA	1452	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG5680-RB	1530	1	100%	0	0%	0	0%	0	0%	0	0%	0	0%	1
CG17684-RA	1546	2	100%	0	0%	0	0%	0	0%	0	0%	0	0%	2
CG11377-RA	1585	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG11125-RA	1591	0	0%	0	0%	0	0%	0	0%	0	0%	1	100%	1
CG11127-RA	1617	1	100%	0	0%	0	0%	0	0%	0	0%	0	0%	1
CG2863-RA	1625	1	100%	0	0%	0	0%	0	0%	0	0%	0	0%	1



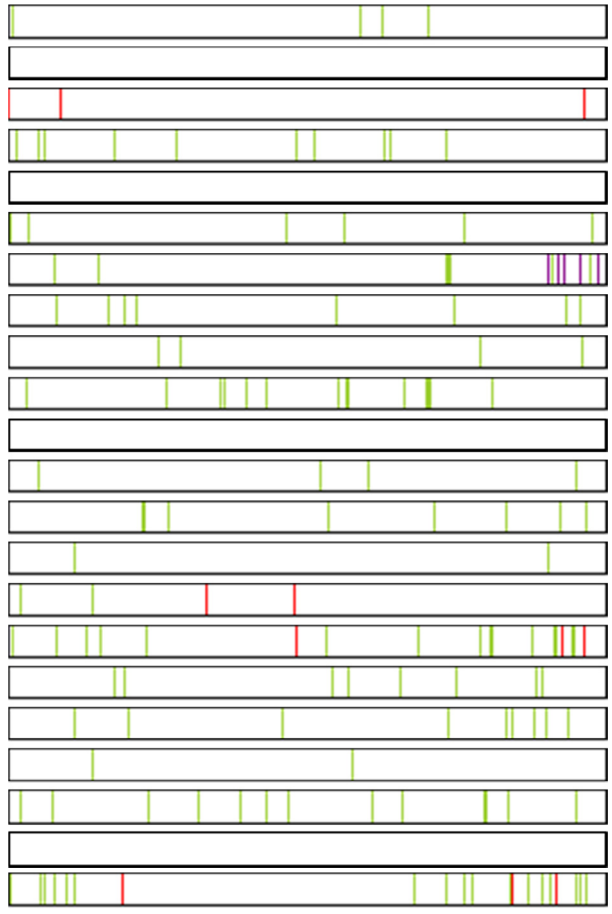
CG1903-RA	1666	1	100%	0	0%	0	0%	0	0%	0	0%	0	0%	1
CG2657-RA	1674	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG40449-RA	1707	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG3709-RA	1719	1	100%	0	0%	0	0%	0	0%	0	0%	0	0%	1
CG31974-RA	1736	1	50%	1	50%	0	0%	0	0%	0	0%	0	0%	2
CG14489-RA	1744	2	67%	0	0%	0	0%	0	0%	0	0%	1	33%	3
CG11023-RA	1773	0	0%	0	0%	0	0%	0	0%	0	0%	1	100%	1
CG2248-RA	1805	2	100%	0	0%	0	0%	0	0%	0	0%	0	0%	2
CG18292-RA	1812	6	100%	0	0%	0	0%	0	0%	0	0%	0	0%	6
CG11372-RA	1832	4	100%	0	0%	0	0%	0	0%	0	0%	0	0%	4
CG5725-RA	1866	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG30110-RA	1908	0	0%	0	0%	0	0%	0	0%	0	0%	1	100%	1
CG32669-RA	1977	0	0%	0	0%	0	0%	0	0%	0	0%	1	100%	1
CG1605-RA	1992	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG2061-RA	1993	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG2720-RA	2000	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG11620-RA	2004	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG11430-RA	2051	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG2522-RA	2074	1	100%	0	0%	0	0%	0	0%	0	0%	0	0%	1
CG11123-RA	2135	5	100%	0	0%	0	0%	0	0%	0	0%	0	0%	5
CG15862-RA	2166	11	100%	0	0%	0	0%	0	0%	0	0%	0	0%	11
CG11450-RA	2185	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0



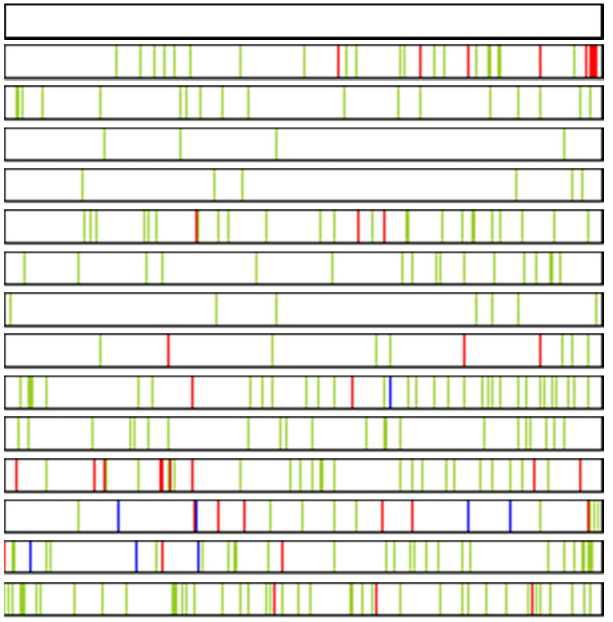
CG1916-RA	2239	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG30325-RA	2253	4	80%	0	0%	0	0%	0	0%	0	0%	1	20%	5
CG2674-RA	2294	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG11186-RA	2299	2	67%	0	0%	0	0%	1	33%	0	0%	0	0%	3
CG11140-RA	2317	1	100%	0	0%	0	0%	0	0%	0	0%	0	0%	1
CG18455-RA	2325	2	100%	0	0%	0	0%	0	0%	0	0%	0	0%	2
CG5779-RA	2344	0	0%	1	50%	0	0%	0	0%	0	0%	1	50%	2
CG3048-RA	2370	1	100%	0	0%	0	0%	0	0%	0	0%	0	0%	1
CG4822-RA	2453	2	100%	0	0%	0	0%	0	0%	0	0%	0	0%	2
CG12017-RA	2466	4	100%	0	0%	0	0%	0	0%	0	0%	0	0%	4
CG8411-RA	2467	2	100%	0	0%	0	0%	0	0%	0	0%	0	0%	2
CG12178-RA	2531	2	100%	0	0%	0	0%	0	0%	0	0%	0	0%	2
CG8390-RA	2539	4	100%	0	0%	0	0%	0	0%	0	0%	0	0%	4
CG5834-RA	2591	9	82%	0	0%	1	9%	1	9%	0	0%	0	0%	11
CG11665-RA	2621	2	100%	0	0%	0	0%	0	0%	0	0%	0	0%	2
CG3164-RA	2768	2	100%	0	0%	0	0%	0	0%	0	0%	0	0%	2
CG5748-RA	2777	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG1616-RA	2801	2	100%	0	0%	0	0%	0	0%	0	0%	0	0%	2
CG1464-RA	2844	1	50%	0	0%	1	50%	0	0%	0	0%	0	0%	2
CG11486-RA	2862	1	100%	0	0%	0	0%	0	0%	0	0%	0	0%	1
CG5036-RA	2869	4	100%	0	0%	0	0%	0	0%	0	0%	0	0%	4
CG4889-RA	2907	2	100%	0	0%	0	0%	0	0%	0	0%	0	0%	2



CG11490-RA	2909	2	100%	0	0%	0	0%	0	0%	0	0%	0	0%	2
CG2144-RA	2914	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG32465-RB	2948	0	0%	1	100%	0	0%	0	0%	0	0%	0	0%	1
CG10283-RA	3001	5	100%	0	0%	0	0%	0	0%	0	0%	0	0%	5
CG4648-RA	3003	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG2331-RA	3018	3	100%	0	0%	0	0%	0	0%	0	0%	0	0%	3
CG4637-RA	3089	5	83%	0	0%	0	0%	1	17%	0	0%	0	0%	6
CG11166-RA	3103	4	100%	0	0%	0	0%	0	0%	0	0%	0	0%	4
CG11121-RA	3104	2	100%	0	0%	0	0%	0	0%	0	0%	0	0%	2
CG15207-RA	3114	9	100%	0	0%	0	0%	0	0%	0	0%	0	0%	9
CG11579-RA	3136	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG9211-RA	3252	2	100%	0	0%	0	0%	0	0%	0	0%	0	0%	2
CG4698-RA	3323	5	100%	0	0%	0	0%	0	0%	0	0%	0	0%	5
CG11066-RA	3391	1	100%	0	0%	0	0%	0	0%	0	0%	0	0%	1
CG18492-RA	3392	1	50%	1	50%	0	0%	0	0%	0	0%	0	0%	2
CG8426-RA	3405	9	90%	1	10%	0	0%	0	0%	0	0%	0	0%	10
CG3836-RA	3637	4	100%	0	0%	0	0%	0	0%	0	0%	0	0%	4
CG3938-RA	3658	5	100%	0	0%	0	0%	0	0%	0	0%	0	0%	5
CG5076-RA	4228	1	100%	0	0%	0	0%	0	0%	0	0%	0	0%	1
CG31973-RA	4324	8	100%	0	0%	0	0%	0	0%	0	0%	0	0%	8
CG2186-RA	4449	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG1624-RA	4516	10	91%	1	9%	0	0%	0	0%	0	0%	0	0%	11



CG10079-RA	4563	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0
CG1708-RA	4772	4	100%	0	0%	0	0%	0	0%	0	0%	0	0%	4
CG42311-RA	4846	13	100%	0	0%	0	0%	0	0%	0	0%	0	0%	13
CG5098-RA	4846	2	100%	0	0%	0	0%	0	0%	0	0%	0	0%	2
CG4952-RA	4943	3	100%	0	0%	0	0%	0	0%	0	0%	0	0%	3
CG40351-RA	5268	13	93%	1	7%	0	0%	0	0%	0	0%	0	0%	14
CG5753-RA	5293	9	100%	0	0%	0	0%	0	0%	0	0%	0	0%	9
CG2671-RA	5407	15	65%	8	35%	0	0%	0	0%	0	0%	0	0%	23
CG2411-RA	5535	5	83%	1	17%	0	0%	0	0%	0	0%	0	0%	6
CG1725-RA	6109	27	90%	2	7%	1	3%	0	0%	0	0%	0	0%	30
CG11376-RA	6463	12	100%	0	0%	0	0%	0	0%	0	0%	0	0%	12
CG2146-RA	6470	15	83%	3	17%	0	0%	0	0%	0	0%	0	0%	18
CG6383-RA	7216	5	63%	2	25%	1	13%	0	0%	0	0%	0	0%	8
CG3936-RA	10189	16	89%	1	6%	1	6%	0	0%	0	0%	0	0%	18
CG9995-RA	11579	25	96%	1	4%	0	0%	0	0%	0	0%	0	0%	26



All 99 cDNA sequences were re-analyzed, now ignoring any predicted off-targets that occur within intron sequences. Although this lowers the predicted off-targets, it still shows a significant occurrence of overlapping off-targets. Our analysis shows that after applying this filter, 74% of the genes show off-targets that occur more than once within the same derived cDNA sequence. The number of items per off-target group is given both in numbers and percentages. Green lines represent sites that share an identical off-target with one other site, red lines represent sites that share an identical off-target with 2 other sites, and blue lines with 3 other sites. Purple lines

represent sites that share identical off-targets with 5 or more other sites. Note that for some genes the lines representing the off-target events are in close proximity and cannot be distinguished as separate lines in this illustrative figure.

Chapter 4

Pantethine rescues a Drosophila model for pantothenate kinase–associated neurodegeneration

Anil Rana, Erwin Seinen, Katarzyna Siudeja, Remco Muntendam, Balaji Srinivasan, Johannes J. van der Want, Susan Hayflick, Dirk-Jan Reijngoud, Oliver Kayser, and Ody C. M. Sibon

Published in Proceedings of the National Academy of Sciences
April 2010

Abstract

Pantothenate kinase–associated neurodegeneration (PKAN), a progressive neurodegenerative disorder, is associated with impairment of pantothenate kinase function. Pantothenate kinase is the first enzyme required for de novo synthesis of CoA, an essential metabolic cofactor. The pathophysiology of PKAN is not understood, and there is no cure to halt or reverse the symptoms of this devastating disease. Recently, we and others presented a PKAN *Drosophila* model, and we demonstrated that impaired function of pantothenate kinase induces a neurodegenerative phenotype and a reduced lifespan. We have explored this *Drosophila* model further and have demonstrated that impairment of pantothenate kinase is associated with decreased levels of CoA, mitochondrial dysfunction, and increased protein oxidation. Furthermore, we searched for compounds that can rescue pertinent phenotypes of the *Drosophila* PKAN model and identified pantethine. Pantethine feeding restores CoA levels, improves mitochondrial function, rescues brain degeneration, enhances locomotor abilities, and increases lifespan. We show evidence for the presence of a de novo CoA biosynthesis pathway in which pantethine is used as a precursor compound. Importantly, this pathway is effective in the presence of disrupted pantothenate kinase function. Our data suggest that pantethine may serve as a starting point to develop a possible treatment for PKAN.

Introduction

CoA is a ubiquitous and essential cofactor for various metabolic reactions, including the tricarboxylic acid cycle and fatty acid metabolism (1). The canonical pathway leading to de novo synthesis of CoA starting at vitamin B5 (also known as pantothenate or pantothenic acid, further referred to as VitB5) is well known. All genes encoding the CoA biosynthetic enzymes have been identified and are highly conserved between different species (1–6) (Fig. 1*A*).

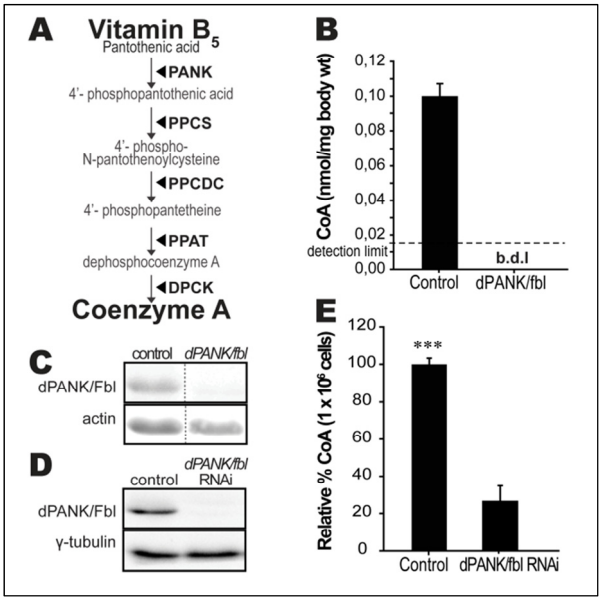


Fig. 1 - *dPANK/Fbl impairment leads to reduced levels of CoA. (A) Scheme of canonical de novo CoA biosynthesis pathway. Vitamin B5 (pantothenic acid) is converted into CoA by the consecutive action of five enzymes: PANK, pantothenate kinase (EC2.7.1.33); PPCS, phosphopantotenoylcysteine synthase (EC6.3.2.5); PPCDC, phospho-N-pantothenoylcysteine decarboxylase (EC4.1.1.36); PPAT, phosphopantetheine adenylyltransferase (EC2.7.7.3); and DPCK, dephospho-CoA kinase (EC2.7.1.24)). (B) HPLC was used to measure levels of CoA in wild-type adult flies and in dPANK/fbl homozygous mutants at 6 days of age. (C) Western blot analysis was used to examine levels of dPANK/Fbl protein in wild types and dPANK/fbl mutants at 6 days of age. Actin was used as a loading control. (D) Western blot analysis was used to measure dPANK/Fbl protein levels in S2 cells 4 days after addition of dPANK/fbl dsRNA. As a control, cells were treated with mock dsRNA.*

*(E) HPLC was used to detect levels of CoA in control Drosophila Schneider's S2 cells and in S2 cells 7 days after dPANK/fbl RNAi treatment. ***P < 0.001 (Student's t test).*

The biosynthesis of CoA, especially the CoA biosynthetic enzyme pantothenate kinase (PANK; EC 2.7.1.33), received renewed interest after the discovery that the Hallervorden-Spatz syndrome, a hereditary disease mainly affecting children, is caused by a mutation in the human *PANK2* gene, one of the four human pantothenate kinase genes (*PANK1-4*), rendering the enzyme inactive (7). Accordingly, this syndrome has been referred to pantothenate kinase–associated neurodegeneration (PKAN). This finding uncovered a completely unknown role of CoA biosynthesis in cellular functioning. Patients with the autosomal recessive disorder PKAN show progressive impairment of speech, locomotor, and cognitive function (8). The pathophysiology of PKAN is not understood, and there is no cure to revert or delay the neurodegeneration. It is not known whether there are decreased levels of CoA in the affected tissues and thus whether decreased levels of CoA coincide with impaired neurological and locomotor function. Although a Pank2 mouse knock-out has been generated, this murine model did not show any signs of neurodegeneration (9), leaving the question unanswered as to whether decreased levels of CoA are causative in PKAN.

Recently, we and others have demonstrated that mutations in *Drosophila* CoA biosynthesis enzymes, including the *Drosophila* homolog of *PANK2* (further referred to as *dPANK/fbl* mutants), induce a neurodegenerative phenotype; and these flies can be used as a model for PKAN-related research (2, 4, 10). *Drosophila* is not only a powerful model to understand the mechanisms of various human neurodegenerative diseases (11), but *Drosophila* disease models are also of value to identify compounds that are able to rescue disease-associated characteristics (12).

In the present study, we used the *Drosophila dPANK/fbl* mutants and dPANK/Fbl down-regulated *Drosophila* cultured S2 cells to address the following questions: (*i*) Does depletion of dPANK/Fbl correlate with decreased levels of CoA? (*ii*) If dPANK/Fbl depletion does induce decreased levels of CoA, are there ways to restore CoA levels in this background? (*iii*) If we are able to restore CoA levels, does this lead to a rescue of the phenotypes induced by dPANK/Fbl depletion?

Our results show that dPANK/Fbl depletion results in a significant decrease of CoA. Furthermore, we tested several compounds for their potential to restore CoA levels in the presence of impaired dPANK/Fbl function. One of the compounds tested was pantethine (the disulphide of pantetheine). Previously, it has been demonstrated that purified enzymatic extracts were able to convert both pantethine and pantetheine into 4′-phosphopantetheine (13–15), an intermediate in the canonical de novo CoA biosynthesis pathway (Fig. 1*A*). However, it has never been tested whether this alternative pathway is functional in a *PANK*-impaired background, although this knowledge is highly relevant in light of a possible PKAN therapy. Feeding pantethine to *dPANK/fbl* mutant flies or adding pantethine to dPANK/Fbl–down-regulated S2 cells restored CoA levels and rescued nearly all tested phenotypes, including the neurodegenerative phenotype. Our data further indicate that pantethine rescued mitochondrial abnormalities in hPANK2-depleted mammalian cells. Our results strongly suggest that pantethine can serve as a precursor compound to generate CoA even in the absence of a functional pantothenate kinase. Our findings may serve as a starting point to develop a possible treatment for PKAN.

Results

dPANK/fbl Mutant Flies Show Reduced Levels of CoA

Pantothenate kinase is the enzyme required for the first step in the canonical biosynthetic route of CoA (Fig. 1*A*). In hypomorphic *dPANK/fbl* mutant flies the dPANK/Fbl protein content was severely decreased (Fig. 1*C*). Although there is a dynamic turnover of CoA in numerous intracellular metabolic reactions, there is only one route known that leads to the de novo synthesis of CoA (1). Therefore, we hypothesized that low levels of CoA caused the phenotype of *dPANK/fbl* mutants. Indeed, HPLC analysis clearly revealed significantly lower levels of CoA in homozygous *dPANK/fbl* mutants compared with wild type (Fig. 1*B*). To further test the effect of impaired function of *dPANK/fbl* on CoA levels, dPANK/Fbl protein levels were down-regulated in *Drosophila* S2 cells by RNAi (experimental setup in Fig. S2). Four days after the addition of *dPANK/fbl* dsRNA, dPANK/Fbl protein levels were strongly decreased (Fig. 1*D*). Under these circumstances, CoA levels were also significantly decreased to 24% of levels of control cells (Fig. 1*E*), and cell counts were significantly lower as

compared with control cells (Fig. 2*A*). This suggested that de novo synthesis of CoA is required for maintenance of normal levels of CoA in *Drosophila* cells and accordingly for normal cell growth in culture.

Pantethine Addition Restores Normal Growth of dPANK/Fbl-Depleted Cells

Our data strongly suggested that reduced levels of CoA might be the primary cause for the decreased cell count of dPANK/Fbl-depleted cells and for the mutant phenotype of *dPANK/fbl* homozygous flies. Accordingly, restoring CoA levels in dPANK/Fbl-depleted cells and in *dPANK/fbl* mutants should lead to a rescue of the related phenotypes. We checked several compounds related to CoA biosynthesis (CoA, VitB5, and pantethine) for their ability to rescue growth of dPANK/Fbl-depleted S2 cells and, when successful, for their ability to restore CoA levels. CoA was tested because adding CoA as a supplement may directly restore CoA levels. VitB5 was tested because adding large doses of VitB5 may compensate for decreased activity of dPANK/Fbl enzyme in a *dPANK/fbl* hypomorphic mutant background. Pantethine was tested because previously it has been reported that pantethine can be converted into the normally occurring CoA intermediate 4′-phosphopantetheine (13–15) (Fig. 1*A*). Our results showed that although high concentrations of all compounds were toxic, CoA and pantethine were effective in restoring cell counts of dPANK/Fbl-depleted cells in a concentration-dependent manner, whereas VitB5 was ineffective (Fig. 2 *B–D*). Because rescue with pantethine was most effective for dPANK/Fbl-depleted cells and pantethine was less toxic compared with CoA, our further analyses were focused on the rescuing potential of pantethine. The optimal effective concentration of pantethine for cells was 100 μM (Fig. 2*D*).

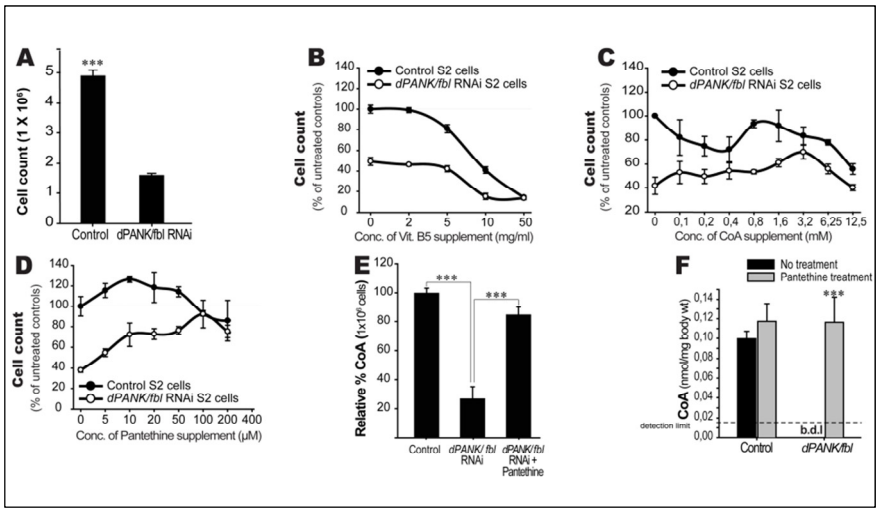


Fig. 2 - *Pantethine rescues cell count of dPANK/Fbl-depleted cells. (A) Control and dPANK/fbl RNAi-treated cells were counted and plated in equal numbers ($0.35 \cdot 10^6$ cells·mL) 4 days after dPANK/fbl RNAi treatment, and proliferation was assayed by counting cells 3 days later. ***P < 0.001 (Student's t test). Error bars indicate SEM. Control cells and dPANK/fbl RNAi-treated cells were plated, and it was tested whether VitB5 (B), CoA (C), or pantethine (D) addition to the medium rescued the cell count of dPANK/Fbl down-regulated cells. 100% represents the number of control cells under normal culturing conditions. (E) CoA levels were measured using HPLC in control cells and in dPANK/fbl RNAi-treated S2 cells (7 days after treatment) with and without addition of 100 µM pantethine. (F) HPLC was used to measure CoA levels in wild types and dPANK/fbl mutants after supplementation of 1.6 mg/mL pantethine to the food. ***P < 0.001 (Student's t test). Error bars indicate SEM. b.d.l, below detection limit.*

Pantethine Rescues Levels of CoA in dPANK/Fbl-Depleted Cells and in *dPANK/fbl* Mutant Flies

First, we tested whether addition of pantethine to the cell culture medium of dPANK/Fbl down-regulated cells could restore CoA levels. Indeed, a restoration of CoA was observed (Fig. 2*E*). Next, we addressed the question whether pantethine addition to the food also rescues *dPANK/fbl* mutants. To identify the effective pantethine concentration in the fly food, various doses of pantethine were tested (Fig. S3). The addition of pantethine at a concentration of 1.6 mg/mL induced a significant increase in climbing

activity in *dPANK/fbl* mutants while inducing only a moderate side effect in wild types. This concentration was used in all further experiments unless indicated otherwise. Concomitantly, levels of CoA were restored upon feeding pantethine directly to *dPANK/fbl* mutants via the food (Fig. 2*F*). These data suggested that there is a dPANK/Fbl-independent way to generate CoA from pantethine in both *Drosophila* S2 cells and in *Drosophila dPANK/fbl* mutant flies. Our data also indicated that pantethine provided via the food can still exert its CoA levels-restoring function.

Mitochondrial Structure and Function Are Severely Affected in *dPANK/fbl* Mutants, and Pantethine Rescues These Phenotypes

Mitochondrial dysfunction is associated with a number of neurodegenerative diseases (16, 17). Several findings suggest that most likely PKAN is also a neurodegenerative disorder associated with impaired mitochondrial function; it has been shown that human PANK2 is localized in mitochondria (18), and that chemical inhibition of pantothenate kinase activity in primary hepatocytes induces abnormal mitochondrial morphology (19). In addition, it was shown that a specific splice isoform of *Drosophila PANK/fbl* was localized in mitochondria of *Drosophila* S2 cells (10). Together these data strongly suggest that human PANK2 and *Drosophila* dPANK/Fbl have a mitochondrial function and that impairment of pantothenate kinase might lead to mitochondrial abnormalities. To investigate this, mitochondrial morphology was examined. Flight muscles contain numerous densely packed mitochondria, and therefore this tissue was analyzed. Visual inspection of flight muscles with bright field microscopy revealed that the structure had a more “loose” appearance and contained more ruptures in *dPANK/fbl* mutants as compared with controls (Fig. 3*A*). Moreover, the muscular degeneration was progressive with age in *dPANK/fbl* mutants. Electron microscopic analysis was performed for a more detailed analysis, and this revealed that, in contrast to wild types, mitochondria of *dPANK/fbl* mutants were severely affected. The mutant mitochondria were swollen and showed damaged cristae and ruptured membranes (Fig. 3*B*). This analysis showed that low levels of CoA coincide with severely damaged mitochondrial structures in *dPANK/fbl* mutants. Pantethine feeding significantly reversed the morphological mitochondrial defects (Fig. 3 *A* and *B*).

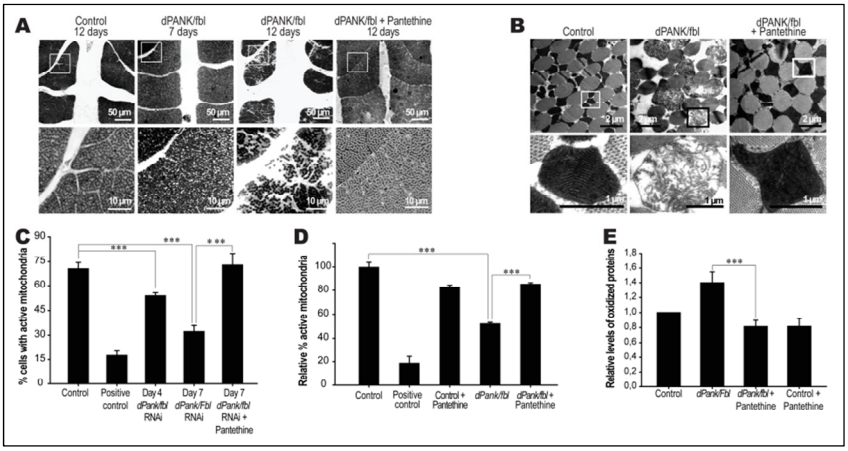


Fig. 3 - *Impaired mitochondrial integrity and increased oxidative damage induced by disruption of dPANK/Fbl function is rescued by pantethine. (A) Morphological analysis of wild-type (12 days old) and dPANK/fblmutant (7 days and 12 days old) flight muscles (untreated and treated with pantethine) was performed by light microscopy. (B) Ultrastructural analysis of mitochondria in flight muscle of 12-day-old wild types and dPANK/fbl mutants (untreated and treated with pantethine). (Lower) Higher magnifications of the indicated areas in Upper. (C and D) JC-1 assay was used to quantify mitochondrial function in control cells and in dPANK/fbl RNAi treated cells (C) in the absence and presence of pantethine and in dPANK/fbl mutants (D) in the absence and presence of pantethine. Valinomycin was used as a positive control. (E) Oxyblots were used to measure levels of oxidative damage to proteins in dPANK/fbl mutants and the effect of supplementation of pantethine was investigated. ***P < 0.001 (Student's t test). Error bars indicate SEM.*

In addition to this analysis, the more quantitative JC-1 assay (Fig. S1 and *SI Text*) was used to measure the percentage of functional mitochondria. Under control conditions, 70.8% of S2 cells had functional mitochondria (Fig. 3*C*). As a positive control for this assay, valinomycin was added to the media, and the percentage of cells with functional mitochondria dropped to 18% (Fig. 3*C*). Under normal culturing conditions, in dPANK/Fbl down-regulated S2 cells, mitochondrial activity was less than 54% after 4 days of RNAi treatment and was less than 32% after 7 days of RNAi treatment compared with control cells (Fig. 3*C*). Mitochondrial function was rescued to the levels of control cells after adding pantethine to the medium of dPANK/Fbl-depleted cells (Fig. 3*C*). A similar assay was performed on isolated mitochondria from *Drosophila flies.* The results showed that *dPANK/fbl* mutants have 42% reduced mitochondrial function at day 6 as compared with control flies (Fig. 3*D*). Interestingly, feeding pantethine to *dPANK/fbl* mutants restored their mitochondrial function up to 84% of wild-type controls (Fig. 3*D*).

Pantethine Reduces Levels of Increased Oxidative Damage of Proteins in *dPANK/fbl* Mutants

Previously, we showed that *Drosophila* CoA mutants displayed an increased sensitivity to oxidative stress (4). Here we investigated whether pantethine was able to reduce increased levels of oxidative stress in *dPANK/fbl* mutants by using Oxyblot assays (Fig. S4). Clearly, *dPANK/fbl* mutants showed increased levels of oxidative damage of proteins compared with wild type, and these levels were strongly reduced by addition of pantethine to the food (Fig. 3*E*).

Pantethine Improves Locomotor Abilities and Rescues Brain Degeneration in *dPANK/fbl* Mutants

So far, our results have demonstrated that pantethine restores CoA levels, improves mitochondrial function and reduces levels of oxidative damage in a *Drosophila* model for PKAN. Next, we investigated whether all these beneficial effects resulted in improvement of locomotor function in *dPANK/fbl* mutants. On average, wild-type larvae were able to crawl 50 cm in 9 min, whereas homozygous *dPANK/fbl* mutants reached only 20 cm (Fig. 4*A*). Addition of pantethine to the larval food significantly improved larval crawling abilities, and an average distance of 30 cm was reached (Fig. 4*A*). Although a strong improvement in larval crawling activity was observed, pantethine feeding was unable to completely rescue the mutant phenotypes. Our data were inconclusive as to whether incomplete rescue was because dPANK/Fbl has additional functions (other than the production of CoA) or whether pantethine has side effects that hampered complete recovery. The latter explanation was supported by our observation that, in wild types, crawling activity was also reduced after pantethine feeding (Fig. 4*A*).

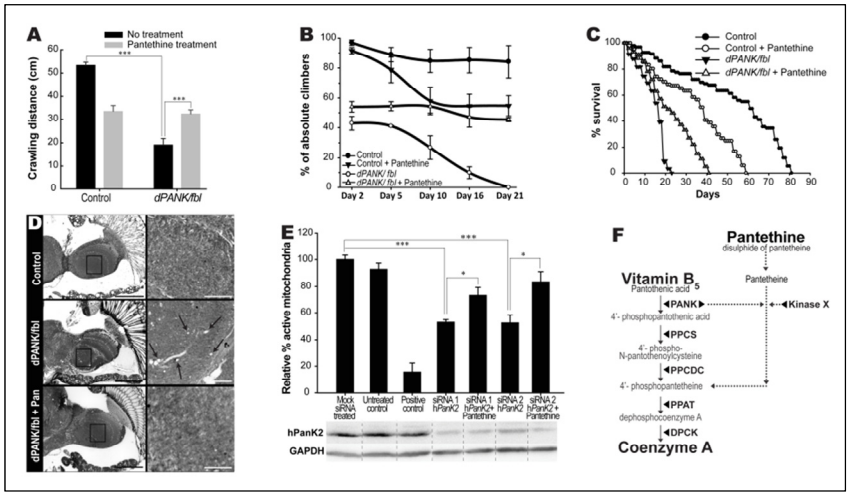


Fig. 4 - *Pantethine is a protective compound in mutant flies and human cells. (A) Larval crawling abilities were measured in wild types and in dPANK/fbl mutants untreated and treated with pantethine. ***P < 0.001 (Student's t test). (B) Percentage of climbers was measured in aging wild types and in aging dPANK/fbl mutants untreated and treated with pantethine. (C) Cohorts of more than 120 flies (n = 3) of wild types and dPANK/fbl mutants were followed and survival curves were generated in the absence and presence of pantethine. All four curves were significantly different compared with each other (log-rank test, P < 0.001). (D) Brain morphology was investigated at the light-microscopic level in (12-day-old) wildtypes, dPANK/fbl mutants untreated and treated with pantethine. (Left) Low magnification. (Bars, 100 μm.) (Right) Higher magnifications of boxed areas. (Bars, 20 μm.) Vacuoles are marked by arrows. (E) Human HEK293 cells were treated with two independent siRNAs (siRNA 1 and siRNA 2) directed against human PANK2 mRNA. Western blot analysis with specific hPANK2 antibodies was used to investigate the effect of the RNAi treatment (48 h after RNAi treatment) on hPANK2 protein levels. GAPDH was used as a loading control. Addition of pantethine (100 μM) to the medium simultaneously with the RNAi treatment resulted in restoration of mitochondria function. Valinomycin was used as a positive control. ***P < 0.001; *P < 0.05 (Student's t test). Error bars indicate SEM. (F) Scheme representing possible pathways for CoA biosynthesis from pantethine. Pantethine may be converted to pantetheine; pantetheine may be phosphorylated by a yet-unknown pantetheine kinase, other than pantothenate kinase (indicated by kinase X) Phosphorylated pantetheine (4'-phosphopantetheine) may enter the canonical de novo CoA biosynthesis pathway downstream of PPCDC and upstream of PPAT.*

Previously, we demonstrated that *dPANK/fbl* flies showed reduced ability to climb at a young age (4). Here we assayed whether this reduced ability to climb further deteriorates with age. Climbing tests of wild-type and homozygous *dPANK/fbl* flies at increasing age (2, 5, 10, 16, and 21 days),

showed that mutants not only possessed impaired climbing abilities following eclosion but that they also experienced a steeper decline of the already reduced climbing activity over time as compared with wild type (Fig. 4*B*). Pantethine feeding significantly prevented the rapid decline of climbing ability of *dPANK/fbl* mutant flies (Fig. 4*B*). Consistent with the data presented in Fig. 4*A*, pantethine feeding induced a decrease in climbing activity in wild-type flies. In addition to these behavioral assays we tested a possible protective function of pantethine on neurodegeneration more directly by analyzing *dPANK/fbl* mutant brain tissue. *dPANK/fbl* mutants show increased numbers of vacuoles in their brains (4), indicating brain degeneration (Fig. 4*D*). Pantethine also rescues this apparent neurodegenerative phenotype (Fig. 4D and Fig. S8).

Pantethine Increases Lifespan of *dPANK/fbl* Mutants

Previously, we have demonstrated that *dPANK/fbl* mutants showed a severe reduction in lifespan (4). We investigated whether rescue of all of the above mentioned phenotypes with pantethine also coincides with increased lifespan of *dPANK/fbl* mutants. The maximal and median lifespan of *dPANK/fbl* mutants were 23 and 15 days, respectively (Fig. 4*C*). Under these circumstances, wild-type flies showed maximal and median lifespans of 81 and 45 days, respectively (Fig. 4*C*). Pantethine feeding increased the maximal lifespan of *dPANK/fbl* mutants from 23 days to 41 days and the median lifespan from 15 to 22 days (Fig. 4*C*). Consistent with the data presented in Fig. 4 *A* and *B,* pantethine feeding induced a reduction in lifespan in wild-type flies. Regardless of these deleterious side effects of pantethine in wild types, pantethine feeding clearly rescued various relevant abnormalities of *dPANK/fbl* mutants.

In Mammalian HEK293 Cells with Down-Regulated PANK2 Levels, Pantethine Also Improved Mitochondrial Function

Finally, we addressed the question of whether pantethine was also capable of rescuing an abnormal phenotype induced by impaired function of endogenous PANK2 in human cells, and for this we used mitochondria integrity as a read-out. First we tested whether down-regulation of PANK2 in HEK293 cells also results in decreased mitochondrial activity [quantified by the mitochondrial JC-1 assay (*SI Text*)]. Indeed, depletion of human

PANK2 using two independent siRNAs resulted in decreased levels of PANK2 and decreased mitochondrial activity (Fig. 4*E*). Addition of pantethine to the medium of PANK2-depleted cells resulted in a significant rescue of mitochondrial activity (Fig. 4*E*). These data indicate that, also in human cells, pantethine was capable of protecting (albeit partly) against consequences of impaired PANK2 enzyme function.

Discussion

In the current study, we used a *Drosophila* model for PKAN to investigate the consequences of impaired pantothenate kinase function and to identify possible protective compounds against the mutant phenotypes. We demonstrate that in *Drosophila dPANK/fbl* mutants and in dPANK/Fbl down-regulated S2 cells, CoA levels are significantly decreased. Low levels of CoA coincide with impaired mitochondrial integrity, increased levels of oxidized proteins, increased loss of locomotor function, neurodegeneration, and decreased lifespan. Our data are consistent with published data demonstrating that numerous neurodegenerative disorders are tightly linked to mitochondrial dysfunction and increased levels of oxidative stress (16, 17). All of the *dPANK/fbl* phenotypes, including neurodegeneration, were more or less rescued by addition of the compound pantethine to the food. Our data support that the mechanism underlying pantethine protection in *dPANK/fbl* flies is specific and not general, because three other neurodegenerative (Parkinson's and two PolyQ) Drosophila models are not rescued by pantethine treatment (Fig. S6). Pantethine has been already proved to be an effective treatment for hyperlipoproteinemia and dyslipidemia in human patients, and a dose of up to 1,200 mg pantethine per day for 24 weeks is tolerated without any side effects (20, 21). Unfortunately it is currently unknown whether pantethine crosses the blood–brain barrier, although this knowledge is highly relevant to develop pantethine further as a possible treatment for PKAN.

For the first time, we show genetic evidence for the existence of a parallel pathway to the canonical de novo CoA biosynthesis starting from pantethine, which at least bypasses the first step of the pathway. We demonstrated that decreased levels of CoA were a clear consequence of impaired dPANK/Fbl function in *Drosophila.* Although this was an anticipated result, this consequence of impaired pantothenate kinase

function has not been investigated in multicellular animals or in human patients. However, there are several reports that indirectly support our observations. Chemical inhibition of PanK1, PanK2, and PanK3 by HoPan in isolated murine hepatocytes resulted in a reduction of de novo CoA synthesis and reduction of total levels of CoA (19). In *Arabidopsis thaliana*, it was demonstrated that mutations in several genes coding CoA biosynthesis enzymes resulted in impaired CoA biosynthesis and decreased levels of CoA (22–24). Together, these and our data show that impaired function of CoA biosynthesis enzymes (including PANK) lead to a decreased rate of de novo CoA synthesis, and that normal de novo synthesis of CoA is required to maintain the physiological levels of CoA.

After establishing that CoA levels were indeed below detection in the *Drosophila* PKAN model, we demonstrated that pantethine is a very potent compound that can act as a starting substrate for generating CoA in a *dPANK/fbl* mutant background. How, exactly, pantethine can be converted into CoA is currently unclear. Classic biochemical studies using cell extracts showed that pantethine can be reduced into pantetheine (25, 26) and that this can be converted into 4′-phosphopantetheine (14). The latter is an intermediate of the canonical de novo biosynthesis pathway and thus here, upstream from PPAT, the CoA de novo synthesis pathways starting from vitB5 and from pantethine may converge (Fig. 4*F*). This is further supported by experiments showing that the decreased cell counts of dPPCS-depleted cells, but not of dPPAT-depleted cells, is rescued by pantethine (Fig. S7). Possible enzymes that catalyze the phosphorylation of pantetheine have never been genetically identified. However, it has been shown that specific purified enzyme preparations were able to phosphorylate both VitB5 and pantethine with similar kinetics (13). Based on these studies it was assumed, but never tested, that pantothenate kinase is the only enzyme present that can phosphorylate pantetheine. However, these earlier studies did not exclude the presence of additional kinases (other than pantothenate kinase) in the purified enzyme preparations with pantetheine kinase activities. It is also possible that pantethine can be converted into CoA not via 4′-phosphopantetheine but via a completely alternative route. Regardless of the exact route, our data suggest that this pathway can most likely occur independently from pantothenate kinase based on the following. (*i*) In both *dPANK/fbl* mutants and in dPANK/fbl down-regulated cells, the levels of dPANK/fbl are severely reduced and it is unlikely that these reduced dPANK/fbl protein levels are responsible for the restoration of CoA levels

after pantethine addition. (*ii*) If some residual pantothenate kinase activity were present in *dPANK/fbl* mutants and in dPANK/Fbl down-regulated cells, addition of extra VitB5 should have been beneficial also. However, addition of extra vitB5 did not lead to rescue in *dPANK/fbl* (Fig. S5) mutants and in dPANK/Fbl down-regulated cells (Fig. 2*B*). Thus all of the above results suggest the presence of an alternative route for de novo synthesis of CoA independent from pantothenate kinase. Apart from whether residual activity of pantothenate kinase is required for pantethine rescue, our findings are still relevant for PKAN-related research because most of the patients with *PANK2* gene mutations still have some residual activity of pantothenate kinase (18).

Regardless of the mechanisms behind pantethine toxicity in wild-type cells, the exact pathway of pantethine conversion into CoA, and whether residual activity of pantothenate kinase is required for pantethine rescue, our data at least suggest the existence of an alternative pathway that uses pantethine as a primary compound to generate de novo CoA in the presence of impaired pantothenate kinase function. This knowledge allows the development of a possible future therapy for PKAN.

Material and Methods

Drosophila Strains

The *Drosophila* strain y^1w^{1118} was used as a wild-type control (Bloomington Stock Centre). The hypomorphic *dPANK/fbl*^{1} mutant flies were used for all assays (2, 4).

Pantethine Supplementation

D-Pantethine (Sigma) was added at a concentration of 1.6 mg/mL in standard fly food; 100 μM of D-Pantethine was added to S2 cell and HEK293 medium, except where mentioned otherwise.

Physiological Assays

To study larval crawling, late third instar homozygous *dPANK/fbl* larvae were placed on 1% nonnutritive agar in a Petri dish. Total distance crawled by the larvae during 9 min was measured. To study the adult lifespan, newly eclosed flies ($n > 100$, 1 or 2 days old) were collected and raised on standard medium at 25 °C in a dry Petri dish with food (2.29 cm^2; with or without pantethine) at the center of the Petri dish. The number of dead flies was counted every 2 days. Each experiment was repeated three times. For a climbing assay, adult flies were used to investigate climbing performance as previously described (27). The experiment was repeated three times ($n > 100$).

CoA Measurements

CoA levels were measured from fly extracts (100 flies, 6 days old) or from *Drosophila* Schneider's S2 cells using HPLC (sample preparation and HPLC analysis are described in *SI Text*).

Cell Culture and PANK Knockdown

Drosophila Schneider's S2 Cells were cultured, and RNAi knockdown of dPANK/Fbl was performed as previously described (28) (*SI Text*).

Mammalian cell culture and siRNA knockdown of hPANK2 are described in *SI Text*.

Electron and Light Microscopy

Flies were immersed in fixative solution (2.5% glutaraldehyde in 0.1 M cacodylate, pH7.8). Postfixation was performed in 2% OsO_4 for 2 h at 4 °C. Dehydration was carried out with graded ethanol series followed by a propylene wash and preembedding in (1:1) propylene:epon solution. Embedding of the flies was performed in EPON. For light microscopy, sections (1–2 μm) were cut using a Reichert Ultracut microtome and stained with Toluidine Blue. For ultrastructural analysis of mitochondria, thin sections (60 nm) were cut from the same samples and analyzed by electron microscopy.

Mitochondrial Assays

Mitochondria were isolated from 7 day old flies as previously described (29). For measurement of mitochondrial membrane potential, J-aggregate-forming lipophilic cation (JC-1) was used to evaluate mitochondrial damage (30). The JC-1 assay (Sigma) was performed according to the manufacturer's manual (*SI Text* and Fig. S1).

Protein Oxidation Detection

Protein lysates were prepared in RIPA buffer containing 2% β-mercaptoethanol. Total protein solutions were incubated with 2,4-dinitrophenylhydrazine (DNP) according to the OxyBlot protein oxidation detection Kit (Chemicon). The total amount of oxidized proteins was quantified for each sample by measuring chemiluminescence from the whole lane and oxidized protein levels were normalized using β-actin as a loading control (Fig. S4).

Antibodies

dPANK/Fbl (1:4,000) (4), hPANK2 (1:2,000; a gift from J. Gitschier, University of California–San Francisco), GAPDH (1:10,000; Fitzgerald Industries), β-actin, and γ-tubulin (Sigma) were used. HRP-conjugated

antimouse or antirabbit antibodies were used (1:2,000; Amersham) as secondary antibodies.

Acknowledgments

We thank Floris Bosveld and Harm Kampinga for stimulating discussions. This work was supported by a VIDI grant (to O.S.), a Neurodegeneration with Brain Iron Accumulation Disorders Association grant (to O.S and S.H.), and a Topmaster grant from the Graduate School GUIDE (to A.R.).

References

1. Leonardi, R., et al., *Coenzyme A: back in action.* Prog Lipid Res, 2005. **44**(2-3): p. 125-53.
2. Afshar, K., et al., *fumble encodes a pantothenate kinase homolog required for proper mitosis and meiosis in Drosophila melanogaster.* Genetics, 2001. **157**(3): p. 1267-76.
3. Begley, T.P., C. Kinsland, and E. Strauss, *The biosynthesis of coenzyme A in bacteria.* Vitam Horm, 2001. **61**: p. 157-71.
4. Bosveld, F., et al., *De novo CoA biosynthesis is required to maintain DNA integrity during development of the Drosophila nervous system.* Hum Mol Genet, 2008. **17**(13): p. 2058-69.
5. Daugherty, M., et al., *Complete reconstitution of the human coenzyme A biosynthetic pathway via comparative genomics.* J Biol Chem, 2002. **277**(24): p. 21431-9.
6. Kupke, T., P. Hernandez-Acosta, and F.A. Culianez-Macia, *4'-phosphopantetheine and coenzyme A biosynthesis in plants.* J Biol Chem, 2003. **278**(40): p. 38229-37.
7. Zhou, B., et al., *A novel pantothenate kinase gene (PANK2) is defective in Hallervorden-Spatz syndrome.* Nat Genet, 2001. **28**(4): p. 345-9.
8. Gregory, A., B.J. Polster, and S.J. Hayflick, *Clinical and genetic delineation of neurodegeneration with brain iron accumulation.* J Med Genet, 2009. **46**(2): p. 73-80.
9. Kuo, Y.M., et al., *Deficiency of pantothenate kinase 2 (Pank2) in mice leads to retinal degeneration and azoospermia.* Hum Mol Genet, 2005. **14**(1): p. 49-57.
10. Wu, Z., et al., *Pantothenate kinase-associated neurodegeneration: insights from a Drosophila model.* Hum Mol Genet, 2009. **18**(19): p. 3659-72.
11. Lessing, D. and N.M. Bonini, *Maintaining the brain: insight into human neurodegeneration from Drosophila melanogaster mutants.* Nat Rev Genet, 2009.
12. Faust, K., et al., *Neuroprotective effects of compounds with antioxidant and anti-inflammatory properties in a Drosophila model of Parkinson's disease.* BMC Neurosci, 2009. **10**: p. 109.
13. Abiko, Y., *Investigations on pantothenic acid and its related compounds. IX. Biochemical studies.4. Separation and substrate specificity of pantothenate kinase and phosphopantothenoylcysteine synthetase.* J Biochem, 1967. **61**(3): p. 290-9.
14. Levintow, L. and G. Novelli, *The synthesis of coenzyme A from panthetheine: preparation and properties of panthetheine kinase.* J Biol Chem, 1954. **207**(2): p. 761-5.

15. Ward, G.B., G.M. Brown, and E.E. Snell, *Phosphorylation of pantothenic acid and pantethine by an enzyme from Proteus morganii.* J Biol Chem, 1955. **213**(2): p. 869-76.
16. Knott, A.B., et al., *Mitochondrial fragmentation in neurodegeneration.* Nat Rev Neurosci, 2008. **9**(7): p. 505-18.
17. Lin, M.T. and M.F. Beal, *Mitochondrial dysfunction and oxidative stress in neurodegenerative diseases.* Nature, 2006. **443**(7113): p. 787-95.
18. Kotzbauer, P.T., et al., *Altered neuronal mitochondrial coenzyme A synthesis in neurodegeneration with brain iron accumulation caused by abnormal processing, stability, and catalytic activity of mutant pantothenate kinase 2.* J Neurosci, 2005. **25**(3): p. 689-98.
19. Zhang, Y.M., et al., *Chemical knockout of pantothenate kinase reveals the metabolic and genetic program responsible for hepatic coenzyme A homeostasis.* Chem Biol, 2007. **14**(3): p. 291-302.
20. Prisco, D., et al., *Effect of oral treatment with pantethine on platelet and plasma phospholipids in IIa hyperlipoproteinemia.* Angiology, 1987. **38**(3): p. 241-7.
21. Bertolini, S., et al., *Lipoprotein changes induced by pantethine in hyperlipoproteinemic patients: adults and children.* Int J Clin Pharmacol Ther Toxicol, 1986. **24**(11): p. 630-7.
22. Rubio, S., et al., *An Arabidopsis mutant impaired in coenzyme A biosynthesis is sugar dependent for seedling establishment.* Plant Physiol, 2006. **140**(3): p. 830-43.
23. Rubio, S., et al., *The coenzyme a biosynthetic enzyme phosphopantetheine adenylyltransferase plays a crucial role in plant growth, salt/osmotic stress resistance, and seed lipid storage.* Plant Physiol, 2008. **148**(1): p. 546-56.
24. Tilton, G.B., et al., *Plant coenzyme A biosynthesis: characterization of two pantothenate kinases from Arabidopsis.* Plant Mol Biol, 2006. **61**(4-5): p. 629-42.
25. Durr, I.F. and N. Cortas, *The reduction of pantethine by an extract of camel intestine.* Biochem J, 1964. **91**(3): p. 460-3.
26. Fisher, D.H. and M.E. Szulc, *Reduction of pantethine in rabbit ocular lens homogenate.* J Pharm Biomed Anal, 1997. **15**(5): p. 653-62.
27. Palladino, M.J., T.J. Hadley, and B. Ganetzky, *Temperature-sensitive paralytic mutants are enriched for those causing neurodegeneration in Drosophila.* Genetics, 2002. **161**(3): p. 1197-208.
28. de Vries, H.I., et al., *Grp/DChk1 is required for G2-M checkpoint activation in Drosophila S2 cells, whereas Dmnk/DChk2 is dispensable.* J Cell Sci, 2005. **118**(Pt 9): p. 1833-42.
29. Schwarze, S.R., R. Weindruch, and J.M. Aiken, *Oxidative stress and aging reduce COX I RNA and cytochrome oxidase activity in Drosophila.* Free Radic Biol Med, 1998. **25**(6): p. 740-7.

30. Smiley, S.T., et al., *Intracellular heterogeneity in mitochondrial membrane potentials revealed by a J-aggregate-forming lipophilic cation JC-1*. Proc Natl Acad Sci U S A, 1991. **88**(9): p. 3671-5.

Supplementary Figures

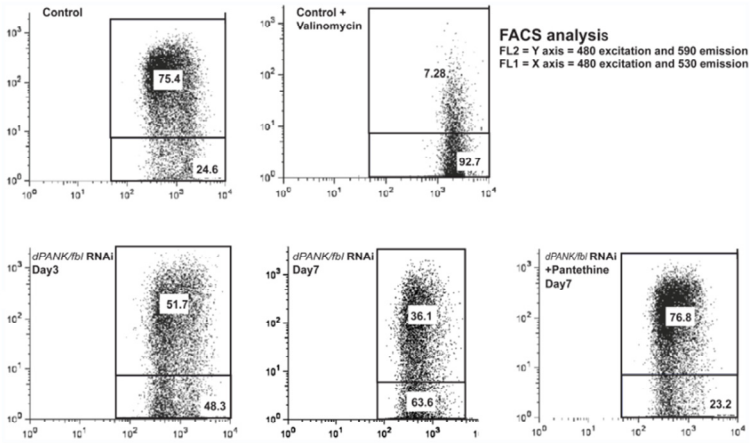


Fig. S1. Analysis of functional mitochondria in dPANK/Fbl depleted cells by FACS analysis. *FACS analysis in combination with a JC-1 assay was used to measure changes in the mitochondria transmembrane potential and this enables the quantification of active mitochondria (see further below for a detailed description of the JC-1 assay). Dot plots are shown for the following conditions: control cells; control cells treated with Valinomycin; dPANK/Fbl depleted cells (untreated and treated with pantethine). The upper boxed areas represent cells with a red and green fluorescence emission of 590 nm and 530 nm above a specific threshold, representing cells with active mitochondria. The lower boxed areas represent cells with a red fluorescence emission of 590 nm below a specific threshold (and with a green fluorescence emission of 530 nm above a specific threshold) representing cells with a disturbed mitochondrial membrane potential. The percentages of cells are indicated in the boxed areas.*

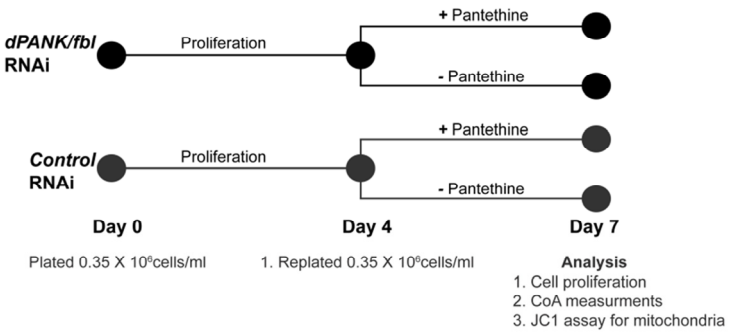


Fig. S2. Scheme of RNAi experiments. *At day 0, cells were plated in equal densities and cells were treated with dPANK/fbl dsRNA or with control dsRNA. After 4 days, cells were replated in equal densities and left further untreated or were treated with pantethine. On day 7 various assays were*

performed. The down-regulation induced by the RNAi treatment for all assays was verified using Western blot analysis using dPANK/Fbl antibodies.

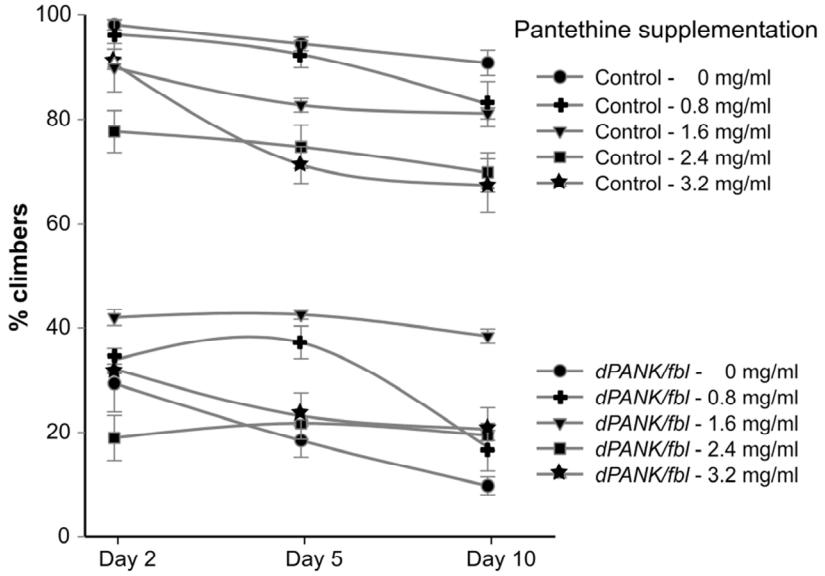


Fig. S3. 1.6 mg pantethine per ml food is the optimal concentration to rescue climbing ability. *Various concentrations (0.8 mg/ml; 1.6 mg/ml; 2.4 mg/ml and 3.2 mg/ml) of pantethine were added to the food of wild type flies and dPANK/fbl mutants. Pantethine was added immediately after eclosion and the food was refreshed every day. On day 2, 5 and 10 climbing activity was measured. Adding 1.6 mg of pantethine per ml of food induced a significant rescue of climbing activity. Moreover, 1.6 mg of pantethine showed only a mildly reduction of climbing activity in wild type flies as compared to 2.4 mg pantethine. Based on these results 1.6 mg pantethine per ml food was used for the experiments in this manuscript.*

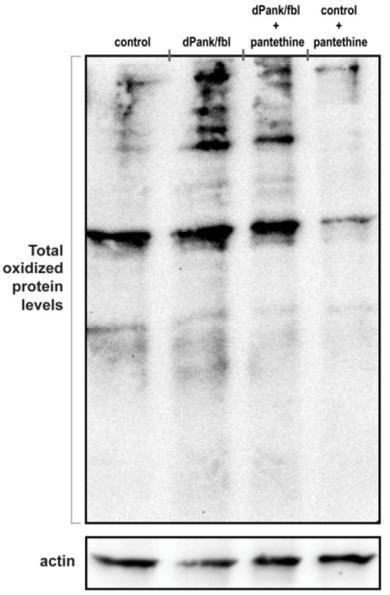


Fig. S4. dPANK/fbl flies have increased levels of oxidized proteins which are rescued upon pantethine feeding. *Oxyblot analysis revealed that dPANK/fbl mutants have higher levels of total oxidized proteins as compared to wild-type. Daily feeding of pantethine for 6 days immediately after eclosion results in reduction of oxidative damage to the proteins in dPANK/fbl mutants. Actin is used as loading control.*

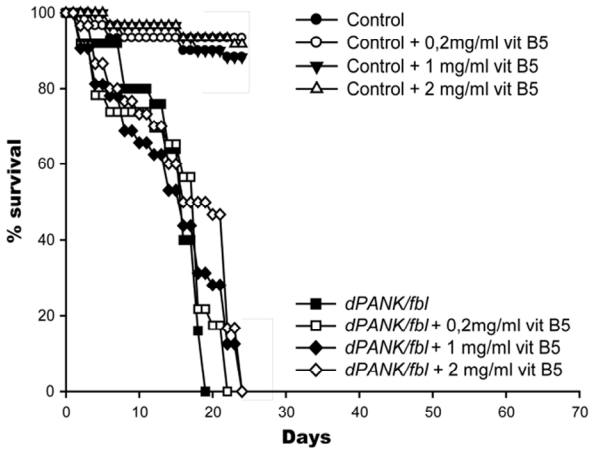


Fig. S5. Vitamin B5 does not increase the life span of dPANK/fbl mutants. *To investigate the effect of vitamin B5, various concentrations (0.2 mg/ml; 1 mg/ml; 2 mg/ml) of vitamin B5 were added to the food and tested for their potential to increase life span. 22 days after feeding the various*

concentrations of vitamin B5, all dPANK/fbl mutants died and no significant increase in life span was observed.

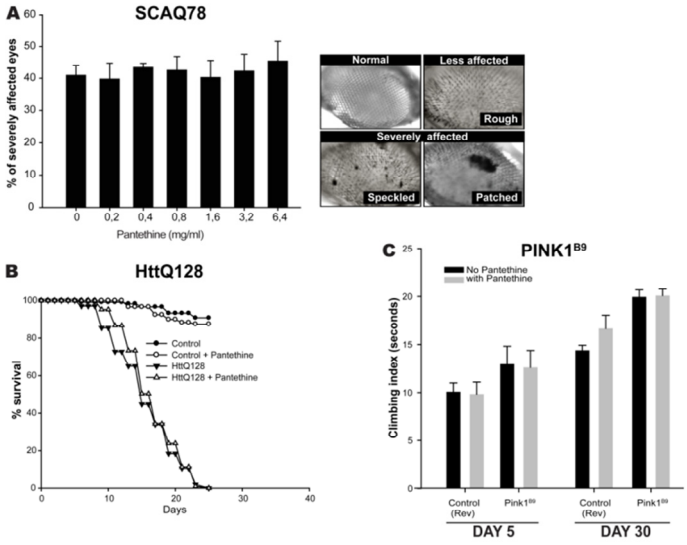


***Fig S6**. Drosophila models for the neurodegenerative diseases: Spinocerebellar Ataxia-type 3 (SCA-3), Huntington and Parkinson are not rescued by pantethine. (A) Flies expressing a truncated form of Ataxin 3 protein containing an expanded repeat of 78 glutamines in eyes show a rough-eye-phenotype [1] and are referred to as SCA3Q78 flies. This transgenic Drosophila strain is a model for SCA3 and has been used to investigate modifiers of toxicity induced by polyglutamine in SCA3 related neurodegeneration. Eye abnormalities are classified as "rough" or as "severely affected" as previously described [2]. Protective compounds will reduce the percentage of severely affected eyes. Increasing concentrations of pantethine did not result in a significant decrease of the percentage of severely affected eyes (>250 eyes were scored for each condition). (B) Flies expressing a truncated form of Huntingtin containing an expanded repeat of 128 glutamines show a neurodegenerative phenotype including a reduced life span of 24 days and are referred to as HttQ128 flies [3]. Addition of (1.6 mg/ml) pantethine to the food did not increase the life span of HttQ128 flies. For each condition over 100 flies were used. (C) Mutations in human PINK1 are linked to parkinsonism. The Drosophila PINK1 gene is an orthologue of the human PINK1 gene and Drosophila PINK1^{B9} mutants show a progressive impairment to climb while they age [4]. Addition of (1.6 mg/ml) pantethine to the food did not improve the climbing ability of PINK1^{B9} mutants. For each time point >100 flies were used. As a control PINK1^{B9} revertants were used that overexpress the wildtype Drosophila PINK1 gene in the PINK1^{B9} mutant background. Climbing index is defined as the average climbing time required to climb 15 cm by 50% of the flies [4].*

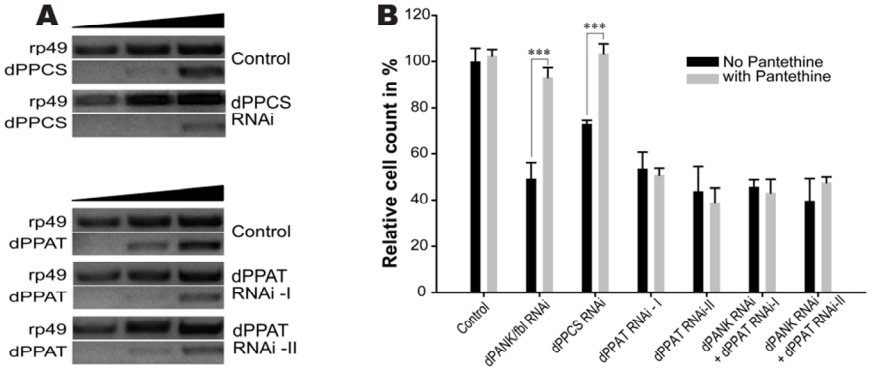


Fig S7. Pantethine rescues the cell count of dPPCS-depleted cells but not of dPPAT-depleted cells. *RNAi was used to down-regulate dPPCS, dPPAT or dPANK/Fbl and dPPAT simultaneously in Drosophila S2 cells. To down-regulate dPPAT, 2 independent non-overlapping RNAi constructs were used (see further below for details of these constructs). Down-regulation induced by the RNAi treatment of dPPCS and dPPAT was investigated by RT-PCR. Down-regulation of dPANK/Fbl induced by RNAi treatment was controlled by using Western blotting (as in Fig 1C). (A) PCR reaction products revealed a significant down-regulation of dPPCS and dPPAT mRNA after RNAi treatment. (B) In dPANK/Fbl-depleted cells, dPPCS-depleted cells, dPPAT-depleted cells and in dPANK/Fbl-dPPAT-double-depleted cells, the cell count was decreased as compared to control cells. Pantethine addition to the cell culture medium significantly increased the cell count of dPANK/Fbl-depleted cells and of dPPCS-depleted cells but not the cell count of dPPAT-depleted cells and of dPANK/Fbl-dPPAT-double-depleted cells. These data strongly suggest that dPPAT is required for the pantethine rescue of dPANK/Fbl-depleted cells. ***, P < 0.001 (Student's t test). Error bars indicate the SEM.*

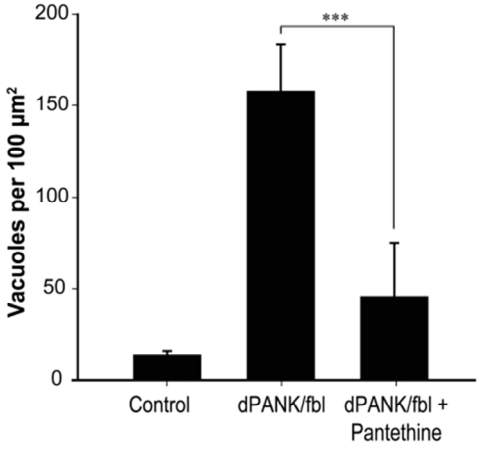


Fig S8. The amount of brain vacuoles in dPANK/fbl mutant flies is decreased after pantethine treatment. *The number of vacuoles in the brain region, indicated in Fig 4D, were measure by NIH Image J software (http://rsb.info.nih.gov/ij/index.html). The method of quantification is outlined in the*

*Image J documentation: "Particle Analysis". The total amount of vacuoles was calculated per 100 μm^2 from comparable regions (indicated in the boxed areas in Fig. 4D). For every condition 4 brains were examined of 12-day-old flies. After eclosion flies were kept on standard food or on standard food supplemented with 1.6 mg/ml pantethine. ***, $P < 0.001$ (Student's t test). Error bars indicate the SEM.*

SI text:

Evaluation of functional mitochondria using a JC-1 assay.

Mitochondrial function can be quantitatively assessed by measuring changes in the mitochondria transmembrane potential using JC-1 which is J-aggregate-forming lipophilic cationic fluorochrome (5,5′,6,6′-Tetrachloro-1,1′,3,3′-tetraethyl-imidacarbocyanine iodide Sigma, USA) assay [5]. At high mitochondrial membrane potentials, JC-1 accumulates in the mitochondria and forms J-aggregates which show a red fluorescence emission at 590 nm. At lower mitochondrial potentials, less dye enters mitochondria resulting in monomers that show green fluorescence emission at 530 nm. Using this assay one can quantify highly active mitochondria (with both red and green fluorescence) and depolarized mitochondria (with green fluorescence only). Using this assay it is possible to investigate the mitochondria integrity of suspension cells, of attached cells and to investigate integrity of isolated mitochondria from tissues. For these assays, valinomycin is added as a control, because this K+ ionophor depolarizes the mitochondrial membrane and is inducing a sharp decrease in red fluorescence emission at 590 nm, representing an increase in impaired mitochondria integrity.

Mitochondrial potential in *Drosophila* S2 cells was estimated using the flow cytometer analysis method for JC-1 probe (Molecular Probes protocol: MitoProbe^{TM} JC-1 assay kit for Flow Cytometry (M34152). Briefly: $1x10^6$ cells were suspended in 500 µl of growth medium and incubated with valinomycin or left untreated. Cells were then centrifuged at 400 rpm (5 min at 4^0C) and resuspended in JC-1 solution buffer (10 µg/ml). After incubation (15 minutes), cells were pelleted, resuspended in ice cold PBS and analyzed immediately using a flow cytometer. Monomers and J-aggregates of JC-1 were simultaneously excited using 488 nm laser and emission was quantified in FL1 (530 nm) and FL2 (590 nm) channels. Mitochondrion containing red JC-1 aggregates (mitochondria with a normal membrane potential; active mitochondria) from viable cells were detectable in FL2 channel, and green JC-1 monomers (mitochondria with a depolarized membrane; impaired mitochondria) were detectable in FL1 channel (Fig. S1). The results were plotted as the percentage of cell with active

mitochondria (FL2) from total number of cells analyzed (20,000 cells per analysis) (Fig. 3C).

Analysis of mitochondria isolated from flies was performed according to the manufacturer's protocol (Sigma, Mitochondrial isolation Kit). In short: 100 µl of the JC-1 Staining Solution was added to 10 µl of isolated mitochondria resuspended in mitochondrial maintenance medium (Sigma, Mitochondrial isolation Kit) in a 96-well plate. Fluorescence was measured in a spectrofluorometer (FL600 Biotek) using the following settings: Excitation wavelength 490 nm; Emission wavelength 590 nm. Fluorescence produced (FLU) per well was recorded and total FLU per mg of proteins (FLU/mgP) was calculated. FLU/mgP is an indication of the amount of J-aggregate formation and a measurement of active mitochondria. In control cells this was set to 100%. The amount of FLU/mgP was indicated for every condition as a percentage of the FLU/mgP in control cells (Fig. 3D).

Analysis of mitochondria in HEK293 cells was performed according to protocols for adherent cells [6, 7]. In short: adherent cells were incubated with JC-1 solution (10 µg/ml) in growth medium for 15 minutes, washed twice with ice cold PBS and fluorescence was measured in a spectrofluorometer. To normalize for the amount of cells, the ratio of FLU for active mitochondria (590 nm alone) to the total FLU from the well (sum of 530 nm and 590 nm) was calculated and in control cells this was set to 100% (Figure 4D).

CoA estimation by HPLC

Fly sample preparation: Homozygous *dPANK/fbl* flies (6 days old), 60 female and 40 males per experiment, were collected and weighed. Flies were then snap frozen in Liquid N_2 and 200µl of solvent buffer (5% sulfosalicylic acid containing 50µM DDT) was added. After thorough grinding, the samples were sonicated 3 times for 10 sec on ice. Samples were centrifuged and supernatant was collected for HPLC analysis of CoA. Prior to analysis, 2µl of Ammonia (25%) was added to 98 µl of the sample solutions. *Drosophila Schneider's S2 cell sample preparation:* Cells were pelleted and 200µl of solvent buffer (5% sulfosalicylic acid containing 50µM DDT) was added. Samples were sonicated, centrifuged and

supernatant was collected for HPLC analysis of CoA. Prior to analysis, 2µl of Ammonia (25%) was added to 98 µl of the sample solutions.

CoA was measured using a slightly modified previously described method [8] using HPLC. A Nucleosil 120 C18 (4.6x150 mm, 3 µm) column was used, together with an Agilent Technologies Guard column C18 (4.6x12.5 mm, 5 µm), with an injection volume of 30 µl per sample. Mobile phase A consisted of 100 mmol/l sodium dihydrogen phosphate and 75 mmol/l sodium acetate. The pH of the buffer was set at 4.6 with phosphoric acid. Mobile phase B consisted of 30% methanol and 70% mobile phase A. The temperature of the column was maintained at 35°C. The solvent gradient consisted of: 10 to 40% B in 10 min, 40% to 90% in 8 min. The column was equilibrated with 10% B between each sample analysis. The flow-rate was maintained at 1.2ml/min. HPLC analysis was performed using a Shimadzu-VP system (Shimadzu ‘s-Hertogenbosch, The Netherlands).

Cell culture and PANK knockdown in *drosophila* and mammalian cells.

Drosophila Schneider's S2 cell culture and RNAi: For generation of the dsRNA the following primers were used:

Gene	Primer
dPANK/Fbl	fwd-CGTGATACGCACCTACAGATG rev-GCCATTGGACCAGAACTCCAT
dPPCS	fwd-GGCACAACAAGCTCCAGAAT rev-CTTGCGTGTCTGCAGCACAT
dPPAT	I) fwd-GCGAGCCATCGAGAAGTACG rev-CCGAGTCATCCAGGAAGATTGT
dPPAT	II) fwd-GCCCACGTGATCGACTGCGAT rev-CCACTTCGCTCAACTTGTTGC

As a control, non-relevant (human gene; hMAZ) dsRNA was used. dsRNA was produced and purified with MEGAscript RNAi Kit (Ambion, USA) according to the manufacturer's instructions. Down regulation of dPANK/Fbl protein was investigated by immunoblotting using dPANK/Fbl specific antibodies for every individual experiment (see also Fig. S2).

Mammalian cell culture and siRNA knockdown of hPANK2: PANK2 knockdown in HEK293 cells was performed under conditions of regulated levels of pantothenic acid (vitB5) using custom made vitB5 free DMEM (ThermoScientific) supplemented with 0,4mg/L vitB5 (Sigma), 10% dialyzed serum (Gibco), 100 U/ml penicillin and 100 µg/ml streptomycin (Invitrogen). Two different human PANK2-specific small interfering RNAs (siRNAs) were used: siRNA1 - Dharmacon (D-003797-04) and siRNA2 - Ambion (AM51321). Nonsilencing control siRNA was purchased from Dharmacon (VOSMC 000005). HEK293 cells were transfected with 100nM siRNA using siPORT Amine transfecting agent (Ambion) according to the manufacturer's protocol. Knockdown efficiency was assayed by immunoblotting 48 hours after transfection.

RT-PCR analysis

Total RNA was extracted from the S2 cells using the Absolutely RNA kit (Stratagene,USA). 1 µg of total RNA was used for cDNA transcription (Invitrogen). Semi-quantitative PCR was performed using the following primers to amplify parts of rp49, dPPCS and dPPAT (product was amplified for 21, 25 and 29 cycles):

Gene	Primer
dPPCS	fwd- ACTTCACCGGCCAGCAGTTC rev- AATCGTCGGCGCTCCATCTC
dPPAT	fwd-GCGAGCCATCGAGAAGTACG rev-CCGAGTCATCCAGGAAGATTGT
rp49	fwd-GCACCAAGCACTTCATCC rev CGATCTCGCCGCAGTAAA

References:

1. Warrick, J.M., et al., *Expanded polyglutamine protein forms nuclear inclusions and causes neural degeneration in Drosophila.* Cell, 1998. **93**(6): p. 939-49.
2. Bilen, J. and N.M. Bonini, *Genome-wide screen for modifiers of ataxin-3 neurodegeneration in Drosophila.* PLoS Genet, 2007. **3**(10): p. 1950-64.
3. Lee, W.C., M. Yoshihara, and J.T. Littleton, *Cytoplasmic aggregates trap polyglutamine-containing proteins and block axonal transport in a Drosophila model of Huntington's disease.* Proc Natl Acad Sci U S A, 2004. **101**(9): p. 3224-9.
4. Park, J., et al., *Mitochondrial dysfunction in Drosophila PINK1 mutants is complemented by parkin.* Nature, 2006. **441**(7097): p. 1157-61.
5. Smiley, S.T., et al., *Intracellular heterogeneity in mitochondrial membrane potentials revealed by a J-aggregate-forming lipophilic cation JC-1.* Proc Natl Acad Sci U S A, 1991. **88**(9): p. 3671-5.
6. Ankarcrona, M., et al., *Glutamate-induced neuronal death: a succession of necrosis or apoptosis depending on mitochondrial function.* Neuron, 1995. **15**(4): p. 961-73.

7. Senoo-Matsuda, N., T. Igaki, and M. Miura, *Bax-like protein Drob-1 protects neurons from expanded polyglutamine-induced toxicity in Drosophila.* Embo J, 2005. **24**(14): p. 2700-13.
8. Demoz, A., et al., *Rapid method for the separation and detection of tissue short-chain coenzyme A esters by reversed-phase high-performance liquid chromatography.* J Chromatogr B Biomed Appl, 1995. **667**(1): p. 148-52.

Chapter 5

A high throughput experimental approach to identify miRNA targets in human cells

Lu Ping Tan, Erwin Seinen, Gerben Duns, Debora de Jong, Ody C.M. Sibon, Sibrand Poppema, Bart-Jan Kroesen, Klaas Kok, Anke van den Berg.

Published in Nucleic Acid Research, September 2009

Abstract

The study of human microRNAs is seriously hampered by the lack of proper tools allowing genome-wide identification of miRNA targets. We performed Ribonucleoprotein ImmunoPrecipitation – gene Chip (RIP-Chip) using antibodies against wild-type human Ago2 in untreated Hodgkin lymphoma (HL) cell lines. 10-30% of the gene transcripts from the genome were enriched in the Ago2-IP fraction of untreated cells, representing the HL miRNA-targetome. In silico analysis indicated that ~40% of these gene transcripts represent targets of the abundantly co-expressed miRNAs. To identify targets of miR-17/20/93/106, RIP-Chip with anti-miR-17/20/93/106 treated cells was performed and 1,189 gene transcripts were identified. These genes were analyzed for miR-17/20/93/106 target sites in the 5' UTRs, coding regions and 3' UTRs. 51% of them had miR-17/20/93/106 target sites in the 3' UTR while 19% of them were predicted miR-17/20/93/106 targets by TargetScan. Luciferase reporter assay confirmed targeting of miR-17/20/93/106 to the 3' UTRs of 8 out of 10 genes. In conclusion, we report a method which can establish the miRNA-targetome in untreated human cells and identify miRNA specific targets in a high throughput manner. This approach is applicable to identify miRNA targets in any human tissue sample or purified cell population in an unbiased and physiologically relevant manner.

Introduction

MicroRNAs (miRNAs) are small RNAs of 19-23 nucleotides which were first discovered less than two decades ago in *Caenorhabditis elegans*(1). Upon binding to Argonaute (Ago) proteins, the RNA induced silencing complex (RISC) is formed for post-transcriptional silencing of genes(2). It is now known that numerous cellular processes including proliferation, differentiation, apoptosis and cell cycle are under regulatory control of miRNAs(3).

Expression of miRNAs can be highly tissue specific(4) and dynamic, as for example seen in hematopoiesis(5;6). The cell physiological impact of miRNA expression was shown by skewing of hematopoeitic stem cell differentiation towards a specific hematopoeitic cell type by changing the expression level of only one miRNA(7). Due to the powerful influence of miRNAs as master regulators of gene expression, it is evident that abnormal expression of miRNAs may contribute to malignant transformation.

Accurate target gene validation has been proven notoriously difficult as apparent by the relatively few miRNA targets that have been experimentally proven thus far. Taken into account that 10-30% of the genes from the genome are predicted to be under the control of miRNAs(8;9), many miRNA:mRNA interactions are still unknown. Several algorithms are available to predict miRNA target genes(8;10;11). However, the consistency between different miRNA prediction algorithms available is limited and the false positive rate is high(8;12). Results from the prediction programs require experimental validation, such as by luciferase reporter assay and Western blotting. Current genome wide screenings approaches include microarray analyses, two-Dimensional fluorescence Difference Gel Electrophoresis (2D-DIGE) and Stable Isotope Labeling with Amino acids in Culture (SILAC)(13;14). However, each of these approaches have their specific caveats including lack of effect at the mRNA level, labor intensiveness, accuracy, complexity of the proteome and protein half-life.

Recently, several studies reported application of an interesting new biochemical approach to analyze cellular mRNA associated with RISC(15-20). In human cells the immunoprecipitation of Ago protein was combined with overexpression of synthetic miRNAs(18-20). Moreover, flag-tagged Ago proteins were used requiring a significant modulation of the cells which

may result in target genes that are not physiologically relevant. The lack of high throughput methods to accurately identify miRNA targets relevant to a specific cell type in an unbiased manner hampers the progression in the discovery of miRNA targets.

In this study, we describe an approach which allows large scale identification of miRNA targets in untreated cells. In this adapted Ribonucleoprotein ImmunoPrecipitation – gene Chip (RIP-Chip) approach, wild-type human Ago2 protein is directly immunoprecipitated from untreated cells. The Ago2-associated mRNA transcripts are analyzed by microarray to identify the miRNA-targetome (whole miRNA regulated gene set) of a specific cell. Moreover, by combining this approach with inhibition of specific miRNAs, we established an approach which allows large scale identification of endogenous transcripts that are targeted by a specific miRNA. This strategy provides unbiased identification of physiologically relevant miRNA target genes.

Materials and Methods

Cell culture and transfection

The HL cell lines, L428 and L1236 were cultured in RPMI 1640 supplemented with ultraglutamine, 100 U/mL penicillin/streptomycin, and 5% or 10% fetal bovine serum (Cambrex Biosciences, Walkersville, USA), respectively. Cells were diluted 1:2 on the day prior to transfection and/or Ago2 immnunoprecipitation.

Locked nucleic acid (LNA) with phosphorothioate (PS) backbone antisense to miR-17-5p, miR-20a, miR-93, miR-106a and miR-106b (Integrated DNA Technologies, Leuven, Belgium) were pooled to form a cocktail of anti-miR-17/20/93/106. LNA antisense to miR-220 was used as a negative control as miR-220 is not expressed in L428(21). Transfection of cell lines was performed using the Amaxa nucleofector I device (Amaxa, Gaithersburg, USA) with solution L, program X-01 for L428 and solution V, program T-01 for L1236. For the RIP-Chip experiment, 5 million cells were transfected with 2.5nmole of anti-miR-17/20/93/106 or anti-miR-220 and the cells were harvested 16h later. Effective silencing of miR-17/20/93/106

in the cell lines (L428 and L1236) had been proven with luciferase reporter assay and Western blot for *CDKN1A*/p21 (data not shown), a proven miR-17 seed family target(22).

RIP-Chip: Immunoprecipitation and western blotting

Immunoprecipitation (IP) of ribonucleoprotein was performed as previously described(23) in both HL cell lines (L1236 and L428) with and without transfection of anti-miR-17/20/93/106 and additionally L428 with transfection of anti-miR-220 as a negative control. Briefly, 10-20 million cells were lysed in 100uL ice cold polysome lysis buffer (5mM MgCl2, 100mM KCl, 10mM Hepes, pH7.0 and 0.5% Nonidet P-40) with freshly added 1mM DTT, 100unit/ml Rnase OUT (Invitrogen, Carlsbad, USA) and 1x complete mini EDTA free protease inhibitor cocktail (Roche, Basel, Switzerland) for 5 minutes. Centrifugation was carried out two times at 14,000g at 4°C for 10 minutes. Supernatant was mixed with 900uL of ice cold NT2 buffer (50mM Tris, pH7.4, 150mM NaCl, 1mM MgCl2, 0.05% Nonidet P-40) containing freshly added 200unit/mL Rnase OUT (Invitrogen, Carlsbad, USA), 0.5% vanadyl ribonucleoside (Invitrogen, Carlsbad, USA), 1mM DTT, 15mM EDTA and 50uL mouse anti-human Ago2 (Clone 2E12-1C9, Abnova, Taipei City, Taiwan) coated sepharose G beads (Abcam, Cambridge, UK). Incubation was carried out overnight at 4°C on a rocking platform. On the following day, beads were washed five times with ice cold NT2 buffer and separated into two portions – one for RNA isolation to identify miRNA target genes and another portion for Western blotting to check for successful immunoprecipitation of Ago2. Mouse IgG_1 isotype control (Abcam, Cambridge, UK) was used as a negative control for the IP procedure. The mouse anti-Ago2 used in the IP was also used for Western blotting at a dilution of 1:1000 while secondary antibody was rabbit anti mouse conjugated with horse radish peroxidase (Dako, Glostrup, Denmark), also at a dilution of 1:1000. For visualization, the blot was incubated 5 minutes with SuperSignal West Pico Chemiluminescent Substrate (ThermoScientific, Rockford, USA) prior to exposure to film.

RIP-Chip: RNA isolation and microarray analysis

RNA from the FT fraction of untransfected L1236, total cell lysate fractions and Ago2-IP fractions of all cells were isolated using Trizol and glycogen (all from Invitrogen, Carlsbad, USA) as a carrier in the ethanol precipitation step. RNA quality was checked with the 2100 Bioanalyzer (Agilent, Santa Clara, USA) and the concentration was determined by Nanodrop 1000 (Thermo Scientific, Wilmington, USA). Microarray analysis was performed according to the manufacturer's protocol (Agilent, Santa Clara, USA). Briefly, first strand cDNA was synthesized from 200ng RNA, followed by cRNA amplification and labeling with Cy3 or Cy5. Purification of Cy3 or Cy5 labeled cRNA was carried out with Qiagen RNeasy Mini kit (Qiagen, Venlo, Netherlands). The cRNA quantity and labeling specificity were determined using the NanoDrop 1000 (Thermo Scientific, Wilmington, USA). Equal amounts of Cy3 or Cy5 labeled cRNA from the Ago2-IP or FT fraction and from the corresponding total cell lysate fraction were mixed and hybridized in a dye swap design at 65°C overnight on Agilent 44k 60mer Human Whole Genome Oligo Microarray. On the following day, slides were washed and signals were scanned with GenePix 4000B (Agilent, Santa Clara, USA). Signal intensities from scanned images were processed and converted into Linear and Lowess normalized data using Agilent Feature Extraction software version 9.1. Quality control report was generated for each array. Using GeneSpring GX version 9.0 (Agilent, Santa Clara, USA), the abundance of probes in the Ago2-IP or FT fraction was compared to the abundance in the total fraction. For each probe, the signal intensity from the Ago2-IP or FT RNA fraction was divided by the signal intensity from the total cell lysate RNA fraction (IP/T or FT/T). For each dye swap, the average IP/T or FT/T ratio was calculated. Any gene transcript corresponding to a probe with an average IP/T ratio of at least 2 was considered as being enriched in the Ago2-IP fraction, representing the miRNA-targetome (Table 1 and supplementary data 1). IP/T of each probe from the untransfected cells was compared to IP/T of those in the cells transfected with antisense oligonucleotides (supplementary data 1). Two criteria were set to identify miR-17 seed family specific miRNA targets: i) the probes must be present in the Ago2-IP fraction (IP/T>2) of untransfected cells and ii) in the transfected cells, the probes should show ≥2 fold depletion from the Ago2-IP fraction when compared to the Ago2-IP fraction of untransfected cells. Different depletion folds from the Ago2-IP fraction upon specific miRNA inhibition were used to generate probe sets for seed

matching site analysis and prediction by algorithms. The data described in this publication have been deposited in NCBI's Gene Expression Omnibus(24) and are accessible through GEO Series accession number GSE14409 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE14409).

In silico analysis of miRNA seed matching sites

For the HL miRNA-targetome, 3' UTRs of all gene transcripts were analyzed for 8mer site matching to the top 5% most abundant miRNAs in L428 and L1236 (23 out of 470 miRNAs assessed, representing 10 miRNA seed families(21), supplementary data 2). MiR-155 was excluded from the search because it was highly expressed in L1236 but only moderately (not top 5%) in L428. To identify targets of the miR-17 seed family, 6mer and 8mer sites in the 3' UTRs were analyzed. The 6mer site is considered as the least stringent requirement for miRNA targeting while an 8mer site is considered to be the most reliable indication for miRNA targeting(14). Additionally, conditions like 5' UTRs, coding sequences and GU wobble were included in the analysis for miR-17 seed family targeting. Since any short sequence has a very high occurrence throughout the genome, the miRNA seed matching sites were first subjected to a background analysis to calculate the percentage of gene transcripts in the genome with the exact site matches within all known 3' UTR sequences of all human genes. The 3' UTR sequences were downloaded from the UCSC website (release March 2006) and loaded into a MySQL database server for convenience using our own programmed importing tool (available upon request). Using specific queries we could identify any miRNA seed matching site and its occurrence throughout the genome. Probe sets generated from the RIP-Chip experiment were analyzed with this method and compared to the results obtained from the genome. All information about miRNA seed matching sites for each probe in the miRNA-targetome can be found in supplementary data 1.

miRNA target prediction

TargetScan release 5.0 was used to predict targets of the 10 miRNA seed familes and miBridge was used to predict miR-17 seed family targets(11;25). For TargetScan, only conserved miRNAs that target conserved gene transcripts were considered. Not all gene transcripts identified in the RIP-

Chip experiment were included in the database of the prediction programs because most of the time only the Refseq transcripts or only one of the 3' UTR for any gene, rather than all 3' UTRs of all transcripts, are included.

Luciferase reporter assay

3' UTRs of 13 genes (Table 2) were cloned into psiCHECK2 vector for luciferase reporter assay (Promega, Madison, USA) as previously described(21). L428 was chosen for luciferase reporter assay because L428 showed higher cell viability and transfection efficiency as compared to L1236. Briefly, the sequence of interest (supplementary data 3) was cloned behind the renilla luciferase (RL) gene in the psiCHECK2 vector. The insert was checked by sequencing. The firefly luciferase (FL) gene present in the same vector was used for normalization to rule out variation in transfection efficiency across samples. 1-2 million L428 cells transfected with 2ug of each construct with or without 2nmole of anti-miR-17/20/93/106 were harvested 48h post transfection for dual luciferase measurement. RL/FL ratio of cells transfected only with the construct was set at 100%. Changes in RL/FL ratio in cells cotransfected with anti-miR-17/20/93/106 are shown as percentage compared to the control. All transfections were repeated at least 3 times to demonstrate consistency of the results and calculate standard deviations.

Results

RIP-Chip of untreated HL cells

Protein coding mRNA transcripts which serve as target genes for miRNAs are bound indirectly to the Ago containing RISC (Fig. 1A). An antibody against wild type human Ago2 was used to immunoprecipitate the RISC from the total cell lysate of HL cell lines L1236 and L428. Immunoprecipitation was validated by Western blotting, showing Ago2 protein in the total cell lysate and the Ago2-IP fraction, while the Ago2 protein was absent in the flow through (FT) fraction. A mouse isotype control (IgG1) antibody was used as a negative control. Here, Western blot showed Ago2 protein in the total cell lysate and in the FT, but not in the IP fraction (Fig. 1B). Using gene expression arrays, the signal intensities of

probes associated with Ago2-IP were compared to the signal intensities of probes in the total cell lysate fraction (IP/T ratio). Different IP/T thresholds (1, 2 and 4) were applied to determine the probes enriched in the IP fraction. A threshold of IP/T>1 resulted in the identification of 12,409 probes in L1236 and 15,178 probes in L428 enriched in the Ago2-IP fractions. An IP/T>2 reduced the number of probes enriched in the IP fraction to 3,164 in L1236 and 2,703 in L428 (Fig. 1C).

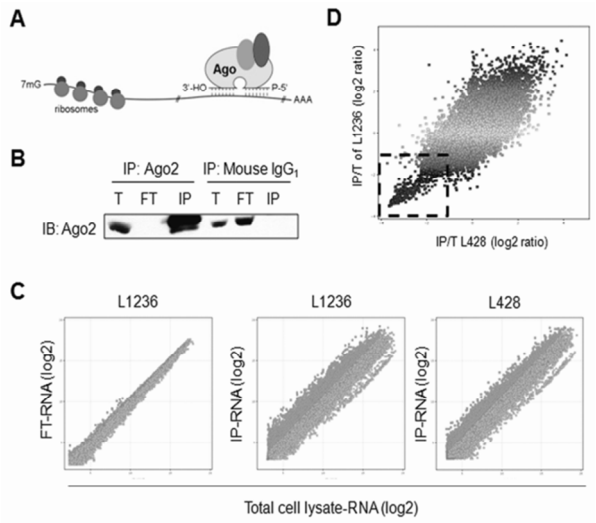


Figure 1. RIP-Chip for identification of Ago2 associated gene transcripts. ***A)*** *Schematic diagram of the RNA induced silencing complex (RISC) mediated gene silencing.* ***B)*** *Immunoprecipitation (IP) of Ago2 complex, analyzed by Western blotting (IB). Ago2 was pulled down when appropriate antibody was used. Mouse IgG1 was used as a negative control and indeed revealed no immunoprecipitation of Ago2.* ***C)*** *Microarray analysis showed that 43 probes in flow through (FT) of L1236, 3,164 probes in IP of L1236 and 2,703 probes in IP of L428 (highlighted in green) were more than 2 fold enriched compared to the total cell lysate (T).* ***D)*** *Probe sets enriched in the Ago2-IP showed a good correlation in both Hodgkin lymphoma cell lines. Probes representing the "non miRNA targets" (IP/T<0.5) were outlined in open box with dash line.*

With a threshold of IP/T>4, the number of probes enriched in the IP fraction was reduced to 882 in L1236 and 398 in L428. As an additional control we also analyzed the FT of L1236 in the same way, this revealed 15,075, 43 and 1 probe using an FT/T threshold of 1, 2 and 4 respectively (Table 1). Based on the results observed in FT/T, and considering the fact that 10-30% of the genes from the genome are predicted to be under the control of miRNAs(8;9), we chose IP/T>2 as the threshold for miRNA target genes and collectively called all gene transcripts in this category as the HL

miRNA-targetome. Interestingly, the miRNA-targetome of L1236 and L428 shared a marked overlap but also showed distinct differences (Fig. 1D). These differences are caused by minor discrepancies in transcriptome and differences in the extent of miRNA regulation between L1236 and L428.

	FT/T	IP/T	
	L1236	**L1236**	**L428**
>4	1	882	398
>2	43	3.164	2.703
>1	15.075	12.409	15.178

Table 1 - *Numbers of probes enriched in FT or Ago2-IP fraction as compared to total cell lysate.*

In silico analysis of the HL miRNA-targetome

3,164 probes (representing 2,746 unique gene transcripts and 2,629 genes) in L1236 and 2,703 probes (representing 2,431 unique gene transcripts and 2,363 genes) in L428 with IP/T>2 were considered as the miRNA-targetome of the corresponding cell lines and these probe sets were studied to assess miRNA targeting *in silico*. The probe set with IP/T<0.5 in both untreated HL cell lines (Fig. 1D) was termed as "non miRNA targets" and used as a negative control. The top 5% expressed miRNAs in L428 and L1236(21) harbored 10 different seed sequences (supplementary data 2). These miRNAs were considered to be the miRNA candidates accountable for a main part of the miRNA-targetome. Two aspects were used as definitions for miRNA targets: 1) presence of miRNA 8mer seed matching site in the 3' UTR and 2) prediction by TargetScan release 5.0. All information about miRNA seed matching sites and TargetScan prediction for each probe in the miRNA-targetome can be found in supplementary data 1.
In the analysis of 3' UTRs for 8mer site matching to each of the 10 seeds of the top 5% expressed miRNAs, the percentage of miRNA targets was always higher in the miRNA-targetome of both cell lines as compared to the genome (Fig. 2). Moreover, about 32% of the miRNA-targetome of both HL cell lines contained the 8mer sites of at least one of the 10 seeds whereas this was only 18% in the genome. Notably, 10% of the miRNA-targetome of both HL cell lines contained at least two 8mer sites of the top 5% expressed miRNAs.

According to the TargetScan release 5.0, the miRNA-targetome of both HL cell lines contained higher percentages of miRNA targets of the 10 seed families when compared to the percentage in the database (Fig. 2). About 30% and 40% of the miRNA-targetome of L1236 and L428 respectively were predicted as targets of at least one of the top 5% expressed miRNAs whereas only 24% of all genes present in the TargetScan database were predicted targets of at least one of the top 5% expressed miRNAs. The percentage of genes predicted to be targeted by at least 2 of the top 5% expressed miRNAs was 9% for all genes in the TargetScan database, 15% and 21% in the miRNA-targetome of L428 and L1236 respectively.

In contrast to the enrichment of miRNA targets observed in the HL miRNA-targetome, the percentage of miRNA targets in “Non miRNA targets” was always lower compared to the genome (Fig. 2).

Our data indicated that we obtained a significant enrichment of miRNA targets in the Ago2-IP fraction (ie miRNA-targetome). In HL, the top 5% expressed miRNAs could already be sufficient to regulate up to about 40% of the miRNA-targetome. The remaining 60% of the miRNA-targetome is assumed to be regulated by other miRNAs which are moderately expressed in HL.

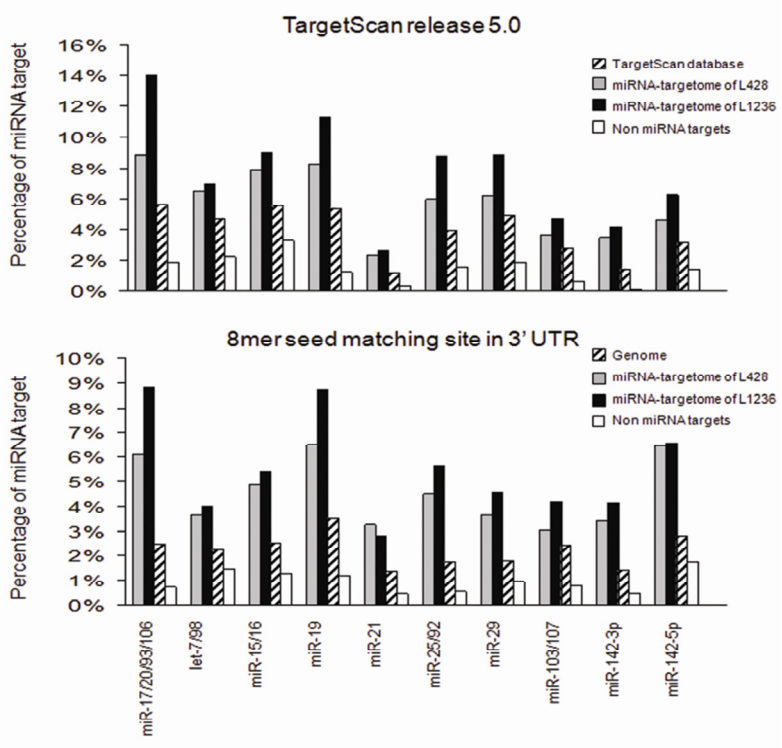


Figure 2. Enrichment of miRNA targets is apparent in miRNA-targetome of HL but not in "non miRNA targets". *Definitions of miRNA targets were 1) prediction of targeting by TargetScan release 5.0 and 2) presence of 8mer site in the 3' UTRs. In both definitions the percentage of miRNA targets was always higher in the miRNA-targetome of HL (L428 and L1236) and lowest in the "non miRNA targets", when compared to genome or the whole database.*

Anti-miRNA strategy combined with RIP-Chip

As miRNAs from the miR-17 seed family comprised a large proportion of the top 5% expressed miRNAs in HL(21) and are frequently associated with the regulation of cell cycle, we proceed with anti-miRNA strategy combined with RIP-Chip to identify the endogenous miR-17 seed family targets in HL.

Upon inhibition of miRNAs of the miR-17 seed family by antisense oligonucleotides, targets of the miR-17 seed family are depleted from the Ago2-IP fraction and remain in the FT fraction (Fig. 3A). In order to identify targets of the miR-17 seed family in a high throughput manner, RIP-Chip was carried out in both HL cell lines (L1236 and L428) with

transfection of anti-miR-17/20/93/106, and the data were compared to the data of untransfected cells. Validation of the IP procedure by Western blots revealed a positive staining of Ago2 in the total and the Ago2-IP fraction whereas no Ago2 was observed in the FT fraction (Fig. 3B). This showed that the immunoprecipitation procedure was successful. RNA was isolated from the total and Ago2-IP fraction of all cells and subjected to microarray analysis. By comparing IP/T values of untransfected and anti-miR-17/20/93/106 transfected cells, 493 probes in L428 and 895 probes in L1236 were ≥2 fold depleted from the Ago2-IP fraction as compared to untransfected cells (highlighted in blue, Fig. 3C). Probes with <2 fold depletion were considered as "not depleted". In addition to the ≥2 fold depletion we also analyzed the probe sets showing ≥3 – 7 fold depletion for targeting by the miR-17 seed family (supplementary data 4). The "not depleted" probes in L428 and L1236 were merged together as the "non miR-17 targets", probe sets with ≥2 depletion fold in both cell lines as the "miR-17 targets", and probe sets with ≥4 fold depletion in L428 and ≥7 fold depletion in L1236 as the "potent miR-17 targets" (gene transcripts with highest depletion fold). Every gene transcript isoform was considered as a unique entity and this led to 3,163 "non miR-17 targets", 1,189 "miR-17 targets" and 66 "potent miR-17 targets". As a control, we also compared the IP/T values of untransfected and anti-miR-220 transfected L428 cells, which revealed only 211 probes that were ≥2 fold depleted from the Ago2-IP fraction (highlighted in blue, Fig. 3C).

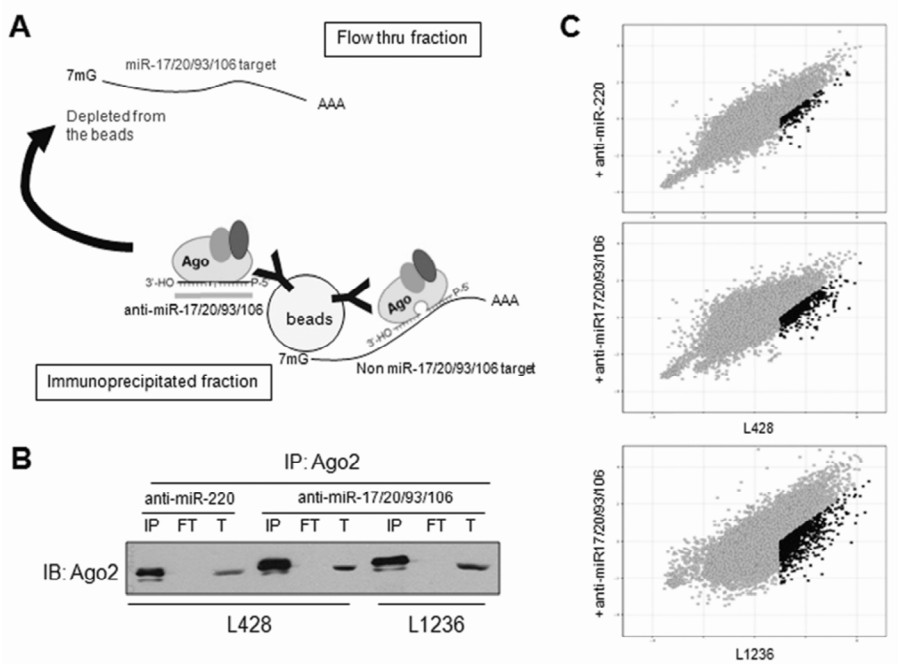


Figure 3. Identification of targets of the miR-17 seed family. A) *Schematic diagram of targets of the miR-17 seed family depleted from the RISC upon miR-17/20/93/106 inhibition. The letter Y represents antibody directed against Ago2 and black thin lines with 7mG in front and AAA in the end represent mRNAs.* ***B)*** *Immunoprecipitation (IP) of Ago2 complex from anti-miR-17/20/93/106 and anti-miR-220 transfected cells, analyzed by western blotting (IB).* ***C)*** *RIP-Chip: Expression profiles of Ago2-IP RNA from anti-miR-220 and miR-17/20/93/106 transfected cells were compared to the untransfected cells. Probes which were present in the untransfected Ago2-IP fraction (IP/T>2) and ≥2 fold depleted from the Ago2-IP fraction upon miRNA inhibition were outlined with dash line.*

Analysis of miR-17 binding sites for miR-17 targets identified in RIP-Chip

Every gene transcript from the human genome was inspected for the presence of miR-17 seed matching sites (6mer and 8mer) in the 3' UTR and analyzed with prediction programs (miBridge and TargetScan). In all analysis, the percentage of putative miR-17 seed family targets showed a gradual increase from the HL miRNA-targetome to "miR-17 targets" and was the highest in the "potent miR-17 targets" (Fig. 4A). 50% of all gene transcripts in "miR-17 targets" and 67% of all gene transcripts in "potent miR-17 targets" contained at least one 6mer matching site in the 3' UTR (Fig.4A). When less stringent conditions (6mer miR-17 seed matching site with GU wobble allowed) were applied and the search was expanded to

include 5' UTRs and coding sequences, 76% of "miR-17 targets" and 82% of "potent miR-17 targets" contained at least one miR-17 seed matching site in anywhere of the entire gene transcript (Fig. 4B, supplementary data 1 and 5).

As a negative control seed matching sites for miR-220 were also analyzed (Fig. 4C) and the percentages were normalized to the percentages found in the genome (Fig. 4D). In this analysis, the enrichment of gene transcripts with miR-17 seed matching site in the "potent miR-17 targets" reached up to 2 fold for 6mer and up to 11 fold for 8mer site while the enrichment of miR-220 seed matching site was consistently low (0-1.3 fold) in all three groups (Fig. 4D).

Higher depletion folds appeared to be correlated with the number and density (average number of 6mer sites/kb) of miR-17 seed matching sites in the 3' UTR (Fig. 5). The "potent miR-17 targets" had the highest density for miR-17 seed matching site compared to "miR-17 targets" and "non miR-17 targets" (Fig. 5A). Enrichment of gene transcripts with multiple miR-17 seed matching sites in the "potent miR-17 targets" reached up to 9 fold for 6mer and 55 fold for 8mer site (Fig. 5B). These results indicated that our approach led to an increased number of gene transcripts with single and multiple miR-17 seed matching sites.

In L428 cells transfected with anti-miR-220, we did not observe a correlation of higher depletion fold with percentage of probes with miR-220 seed matching sites (both 6mer and 8mer) (supplementary data 4). This indicated that no miR-220 targets are identified with this approach, which is consistent with the lack of miR-220 expression in the HL cells. To determine depletion of non-specific probes due to the transfection procedure, we compared anti-miR-220 depleted probes to anti-miR-17/20/93/106 depleted probes. In L428 cells, 154 probes were depleted in both anti-miR-17/20/93/106 and anti-miR-220 transfected cells (supplementary data 1). 62% (96/154) of these consistently depleted probes contain at least one perfect 6mer seed matching site for miR-17 in the 5' UTR, coding region and/or 3' UTR of the gene transcript while 39% of them (60/154) contain at least one perfect 6mer seed matching site for both miR-17 and miR-220 in the 5' UTR, coding region and/or 3' UTR (supplementary data 1). Using less stringent conditions for miR-17 seed family target (ie 6mer seed matching site with GU wobble allowed in 5' UTR, coding region and 3'UTR), 73%

(113/154) of these consistently depleted probes were identified as potential targets of miR-17 seed family. Based on these data we did not exclude them from the "miR-17 target" list.

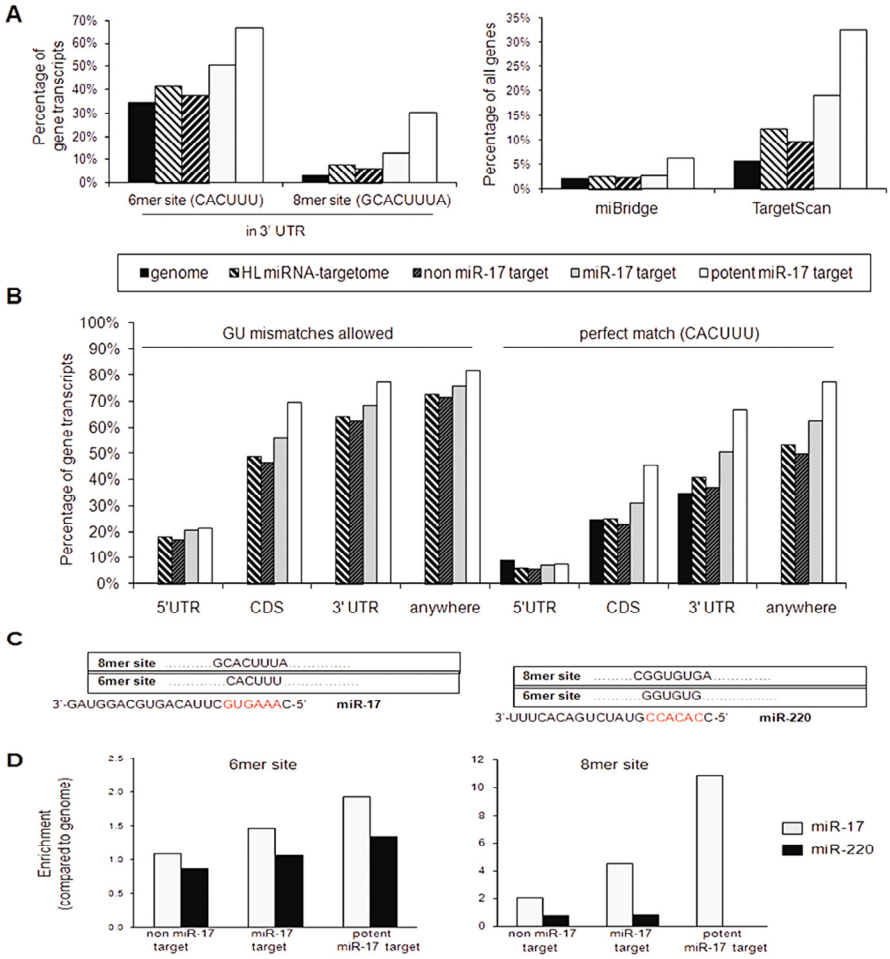


Figure 4. Enrichment of miR-17 seed family targets was correlated with higher depletion fold upon inhibition of the miR-17 seed family*. A) Percentage of miR-17 seed family targets by all definitions (presence of 6mer site and 8mer sites in 3' UTRs, prediction by programs) was always highest in the "potent miR-17 targets". B) Analysis of 6mer miR-17 seed matching site with GU wobble allowed in the entire gene transcript revealed up to 76% of "miR-17 targets" and 82% of "potent miR-17 targets" with at least one miR-17 seed matching site in the entire gene transcript. C) 6mer and 8mer seed matching sites for miR-17 and miR-220. D) Enrichment of gene transcripts with miR-17 seed matching sites is correlated with higher depletion fold upon inhibition of the miR-17 seed family. In contrast, enrichment of miR-220 seed matching site, a miRNA which is not expressed in L428 is minimal.*

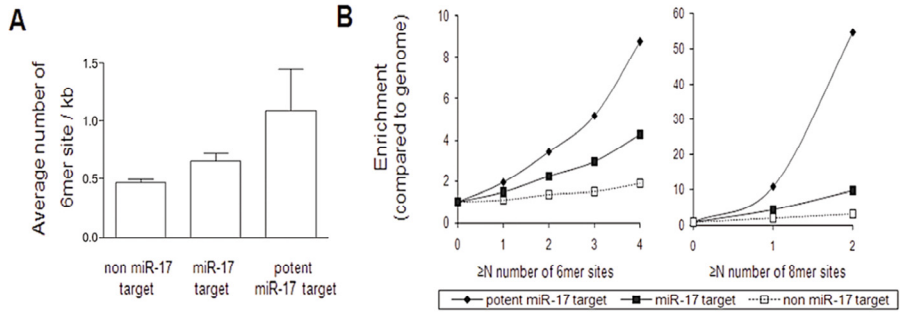


Figure 5. Gene transcripts with higher depletion fold upon inhibition of the miR-17 seed family had higher density and number of miR-17 seed matching sites in the 3' UTRs*. A) "Potent miR-17 targets" had the highest density of miR-17 seed matching sites in the 3' UTRs. Mean with 95% confidence interval is shown. The data is statistically significant, with p<0.0001 (One-way ANOVA). B) Enrichment of gene transcripts with multiple miR-17 seed matching sites was correlated with higher depletion fold upon inhibition of the miR-17 seed family.*

Validation by luciferase reporter assay

To validate the results generated from this RIP-Chip approach, 13 genes, namely *ADRB2, CCL1, CD274, FBXO31, GPR137B, NPAT, OBFC2A, RAB12, RBJ, SGK1, YES1, ZNF22,* and *ZNFX1* (Fig. 6), were chosen for luciferase reporter assay. These genes can be further categorized into groups according to their depletion fold in the Ago2-IP fraction upon miR-17/20/93/106 inhibition ("non miR-17 targets", "miR-17 targets" or "potent miR-17 targets"), presence of 6mer and 8mer site for miR-17 in their 3' UTRs (Table 2).

ADRB2 and *SGK1* were "non miR-17 targets" according to the RIP-Chip experiment and contained no 6mer site for miR-17 in the 3' UTRs. These two genes were negative in the luciferase reporter assay (Fig. 6). *RAB12* was marginally depleted in RIP-Chip (1.73 fold) and was listed as a "non miR-17 target". In contrast to *ADRB2* and *SGK1*, *RAB12* contained a 6mer site for miR-17 and showed enhanced luciferase activities upon inhibition of the miR-17 seed family (Fig. 6). All "miR-17 targets" and "potent miR-17 targets" identified from the RIP-Chip experiment that contained at least one 6mer site for miR-17 in the 3' UTRs (*CCL1, FBXO31, GPR137B, NPAT, OBFC2A, RBJ, YES1,* and *ZNFX1*) showed increased luciferase signals upon inhibition of the miR-17 seed family (Fig. 6). A more pronounced

increase was observed with the "potent miR-17 targets" that contained at least one 6mer site for miR-17. The two "potent miR-17 targets" without a 6mer site for miR-17 in the 3' UTR (*CD274* and *ZNF22*) did not yield increased signals in the luciferase reporter assays upon inhibition of the miR-17 seed family (Fig. 6). Notably, five of the nine genes showing enhanced luciferase signals have at least one 6mer but not a 8mer site for miR-17 in the 3' UTRs (Table 2).

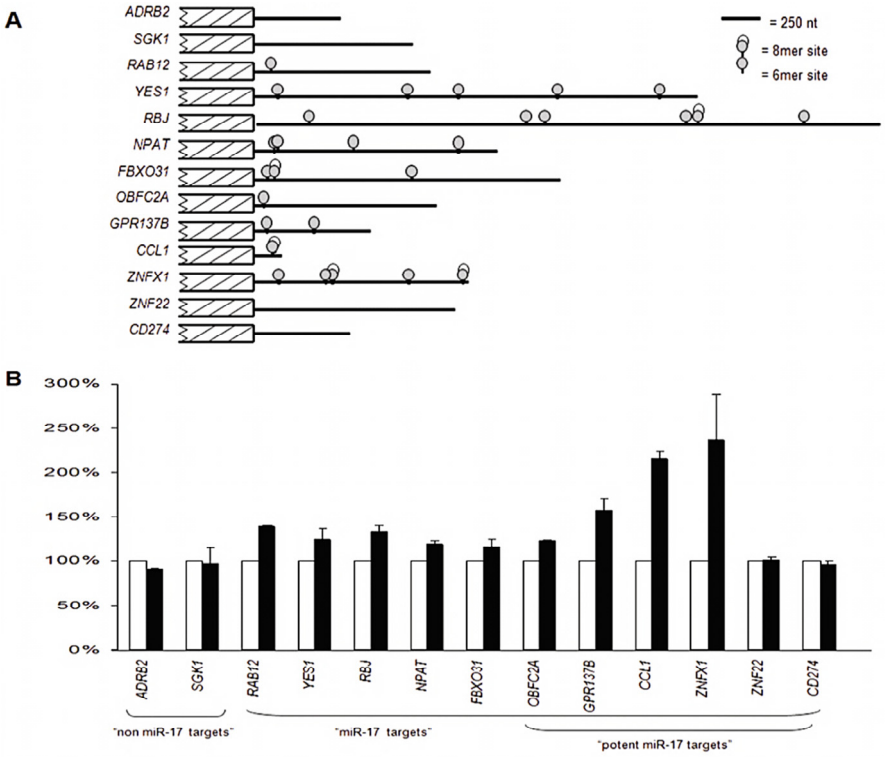


Figure 6. Validation of targets of the miR-17 seed family by luciferase reporter assay. A) *Schematic diagram of the genes cloned into psiCHECK2 vector for luciferase reporter assay. Potential miR-17/20/93/106 binding sites (6mer and 8mer sites) are indicated.* ***B)*** *Luciferase reporter assay for the selected genes confirmed RAB12, YES1, RBJ, NPAT, FBXO31, OBFC2A, GPR137B, CCL1 and ZNFX1 as targets of the miR-17 seed family. Open bar, transfection with construct only. Filled bar, construct co-transfected with anti-miR-17/20/93/106.*

			miR-17 site in 3'		Confirmed by luciferase reporter assay
		Depletion	6mer	8mer	
"non miR-17 targets"					
ADRB2	NM_000024	1.37	0	0	N
SGK1	NM_005627	1.44	0	0	N
RAB12	NM_001025300	1.73	1	0	Y
"miR-17 targets"					
YES1	NM_005433	2.08	5	0	Y
RBJ	NM_016544	2.53	6	1	Y
NPAT	NM_002519	3.54	4	0	Y
FBXO31	NM_024735	3.78	3	1	Y
OBFC2A	NM_001031716	4.64	1	0	Y
GPR137	NM_003272	4.88	2	0	Y
CCL1*	NM_002981	5.67	1	1	Y
ZNFX1*	NM_021035	9.97	5	2	Y
ZNF22*	NM_006963	5.14	0	0	N
CD274*	ENST0000038157	5.23	0	0	N

Table 2. Genes selected for luciferase reporter assay
**also belong to the "potent miR-17 targets". Y, confirmed and N, not confirmed by luciferase reporter assay.*

Discussion

We have demonstrated the effectiveness of a high throughput method for identification of endogenous miRNA targets in untreated human cells. This approach not only allows the analysis of the complete transcriptome for miRNA targets but also permits a more direct identification of physiologically relevant miRNA targets in human cells and tissues. Subsequently, in combination with anti-miRNA strategy the RIP-Chip approach led to high throughput identification of endogenous targets of the miR-17 seed family.

Prediction programs for miRNA targets often predict all possible targets irrespective of their physiologically relevance and their co-expression with the corresponding miRNA. Consequently, the false positive rate for

prediction programs can be high and selection of the most relevant genes from a long list of predicted miRNA targets is difficult. Up to date there are several publications showing the feasibility of the biochemical RISC-IP approach to identify miRNA targets. The experiments described in these publications were performed in *Drosophila melanogaster*(15), by immunoprecipitation of AIN-1 and AIN-2, other RISC associated proteins in *Caenorhabditis elegans*(16), using cloning based strategy in the human embryonic kidney HEK293 cell line(17) and tagged Ago proteins also in the HEK293 cell line(18-20). In the latter studies, experiments were performed using tagged Ago proteins and miRNAs that are not endogenously expressed in HEK293 for miRNA target identification. This may result in targets that normally are not co-expressed with their targeting miRNA and hence the physiological relevance is questionable. The approach we demonstrated here identifies mRNAs which are associated with endogenous miRNA in wild-type human Ago2 containing complex and thus allows direct screening of any human tissue or cell type. Also, we showed that cross-analysis of the results from an anti-miRNA strategy combined with RIP-Chip and presence of a 6mer site in the 3' UTR, irrespective of program prediction, can be sufficient to confirm specific miRNA targeting.

miRNA	Target
let-7/miR-98	*KRAS, CASP3*
miR-15/16	*DMTF1, CCND1, CCNE1*
miR-17/20/93/106	*NCOA3*, RB1*, TGFBR2*, E2F3, ARID4B*, MYLIP, CDKN1A, TP53INP1**
miR-21	*TPM1*
miR-29	*MCL1*

Table 3 . Known miRNA target genes that have been found in the HL miRNA-targetome
**Proven miR-17 targets which were also revealed in the RIP-Chip approach with inhibition of the miR-17 seed family.*

15 known targets of the top 5% expressed miRNAs(13;22;26-36) were identified in our HL miRNA-targetome (Table 3), showing the effectiveness of our approach. Within the HL miRNA-targetome we found a significant enrichment of genes that are associated with the p53 signaling pathway, ubiquitin mediated proteolysis, apoptosis and regulation of cell size (data

not shown), features which are related to the nature of the tumor cells of HL. This HL miRNA-targetome includes genes which are known to be inactivated by mutations in HL cases, like *FAS, NFKB1A, NFKB1E, SOCS1* and *TNFAIP3*(37-42). These results reflect the physiological relevance of our study.

Combining anti-miRNA strategy with RIP-Chip revealed 1,189 gene transcripts ("miR-17 target") that were ≥2 fold depleted from the Ago2-IP upon miR-17/20/93/106 inhibition. Comparison of these 1,189 "miR-17 targets" to the 990 miR-17 target genes predicted by TargetScan (release 5.0) revealed an overlap of about 20%. The limited overlap may be due to the inclusion of all genes with conserved target sites in 3' UTR by the TargetScan prediction program whereas in our experimental approach, we evaluated only endogenous transcripts. According to the results from the luciferase reporter assay, all genes with ≥2 depletion fold and presence of the 6mer site for miR-17 in the 3' UTR can be directly considered as targets of the miR-17 seed family. This resulted in 599 gene transcripts that were identified as miR-17 targets in our approach. However, the remaining 590 gene transcripts included in the "miR-17 targets" list lacked 6mer sites in the 3' UTRs, like *ZNF22* and *CD274*. It might be speculated that the target sites for these genes are present in the coding region and/or the 5' UTRs, as has been reported for p16 regulation by miR-24(43) and *SEC24D* regulation by miR-605(25). To address this question, we re-analyzed the entire miRNA-targetome of HL for presence of miR-17 binding site with the less stringent conditions ie 6mer miR-17 seed matching site with GU wobble allowed. Moreover we expanded the seed matching search into 5' UTRs and coding regions (Fig. 4B, supplementary data 1 and 5). In L428, up to 88% of all "potent miR-17 targets" (including *ZNF22* and *CD274*) and 76% of "miR-17 targets" contained at least one 6mer miR-17 seed matching site (supplementary data 1 and 5). This result is in line with the expectation that miRNA targets should contain target sites matching to the seed sequence of miRNA. In our opinion, the percentage did not reach 100% because the analysis was made with the assumption that only site matching to the 5' seed of the miRNA is important for targeting and sequence complementarity between the 3' end of the miRNA to the target is ignored. We cannot exclude presence of genes in the miRNA-targetome due to non-specific binding. Also, we cannot exclude the possibility of newly acquired or lost of miR-17 seed family target sites in gene transcripts that were expressed in the HL cell models. These two issues can be addressed by application of the

recently published technique called high throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP)(44;45). In HITS-CLIP, the extra crosslinking step induced a covalent binding of the RNA to RISC, allowing a more stringent purification. Also, the exact miRNA binding sites can be identified by deep sequencing.

In the HL miRNA-targetome, co-regulation of a gene transcript by multiple miRNAs is common (as indicated in our *in silico* analysis, supplementary data 1). In the RIP-Chip experiment with inhibition of the miR-17 seed family, we identified five out of eight of the miR-17 targets, *NCOA3, RB1, TGFBR2, ARID4B, TP53INP1* that were experimentally proven elsewhere(26;30-32). The other three experimentally proven miR-17 targets (*E2F3, MYLIP* and *CDKN1A*)(22;28;32) had depletion fold of 1-1.8 and hence were categorized in the "non miR-17 targets". Similarly, we validated *RAB12* as target of the miR-17 seed family but it was only 1.73 fold depleted upon inhibition of the miR-17 seed family. It can be speculated that inhibition of the miR-17 seed family alone is insufficient to remove *RAB12, E2F3, MYLIP* and *CDKN1A* from the Ago2-IP fraction, based on the presence of predicted target sites for let-7, miR-15, miR-25, miR-19 and miR-29 in the 3' UTR of these genes. Although we cannot exclude the presence of miR-17 seed family targets in the "non miR-17 targets" group, these genes most likely are simultaneously targeted by miRNAs other than the miR-17 seed family and hence minimal effects are seen upon inhibition of only miR-17/20/93/106. The complexity of miRNA:mRNA interaction(46) still awaits to be addressed.

Consistent with the observation from the Bartel group who used a proteomics approach to identify miRNA targets(14), we found a higher enrichment of 8mer sites as compared to 6mer sites in the RIP-Chip identified miRNA targets. However, 5 out of 9 targets of the miR-17 seed family verified in our luciferase reporter assay have 6mer but not 8mer site in the 3' UTRs. Our results suggest that presence of the 8mer site in the 3' UTR is a good indicator to identify miRNA-specific targets, but presence of the 8mer site is not obligatory for effective targeting by miRNA. Recently, the Ambros group analyzed mRNA transcripts which were identified by immunoprecipitation of AIN proteins in *C. elegans* and created a program called mirWIP(47). This program considers various features of the experimentally identified miRNA targets (like structural accessibility of target sequences, total free energy of miRNA-target binding) and hence a

better refinement of the miRNA prediction algorithm can be achieved. Within the scope of our study, perfect 6mer seed matching sites in the 3' UTR is still the best criterion to identify miR-17 seed family targets from the genome, as this criterion clearly discriminate the two groups better than other criteria (Fig. 4B, supplementary data 5). Despite the higher coverage observed for the analysis which allowed GU mismatches, these criteria suffer from a high background precluding an effective analysis for enrichment of miRNA targets (Fig. 4B, supplementary data 5). Hence, it will be interesting to include the human miRNA targets identified in our study for mirWIP analysis and sort out the conditions which discriminate between the "miR-17 targets" and background.

In conclusion, we have established a high throughput approach to identify endogenous miRNA targets of untreated human cells and provide an option to evaluate miRNA seed family specific targets. This is an important improvement as current methods lack the advantage of high throughput and unbiased identification of physiologically relevant target genes.

Fundings

This work was supported by the Ubbo Emmius Foundation, University Medical Center Groningen [to L.P. Tan]; the Dutch Cancer Society [2006-3643 to A. van den Berg] and the Netherlands Organization for Scientific Research [NWO VIDI 971-36-400 to O.C.M. Sibon].

Acknowledgments

We thank George Bell from Bioinformatics and Research Computing, Whitehead Institute, for information about the TargetScan 3' UTR database and Inhan Lee from Center for Computational Medicine and Biology, University of Michigan for information about the miBridge database.

Reference List

1. Lee,R.C., Feinbaum,R.L. and Ambros,V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843-854.
2. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281-297.
3. Ambros,V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350-355.
4. Liang,Y., Ridzon,D., Wong,L. and Chen,C. (2007) Characterization of microRNA expression profiles in normal human tissues. *BMC. Genomics*, **8**, 166.
5. Ramkissoon,S.H., Mainwaring,L.A., Ogasawara,Y., Keyvanfar,K., McCoy,J.P., Jr., Sloand,E.M., Kajigaya,S. and Young,N.S. (2006) Hematopoietic-specific microRNA expression in human cells. *Leuk. Res.*, **30**, 643-647.
6. Georgantas,R.W., III, Hildreth,R., Morisot,S., Alder,J., Liu,C.G., Heimfeld,S., Calin,G.A., Croce,C.M. and Civin,C.I. (2007) CD34+ hematopoietic stem-progenitor cell microRNA expression and function: a circuit diagram of differentiation control. *Proc. Natl. Acad. Sci. U. S. A*, **104**, 2750-2755.
7. Chen,C.Z., Li,L., Lodish,H.F. and Bartel,D.P. (2004) MicroRNAs modulate hematopoietic lineage differentiation. *Science*, **303**, 83-86.
8. John,B., Enright,A.J., Aravin,A., Tuschl,T., Sander,C. and Marks,D.S. (2004) Human MicroRNA targets. *PLoS. Biol.*, **2**, e363.
9. Lewis,B.P., Shih,I.H., Jones-Rhoades,M.W., Bartel,D.P. and Burge,C.B. (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787-798.
10. Griffiths-Jones,S., Saini,H.K., van,D.S. and Enright,A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154-D158.
11. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15-20.
12. Sethupathy,P., Megraw,M. and Hatzigeorgiou,A.G. (2006) A guide through present computational approaches for the identification of mammalian microRNA targets. *Nat. Methods*, **3**, 881-886.

13. Zhu,S., Si,M.L., Wu,H. and Mo,Y.Y. (2007) MicroRNA-21 targets the tumor suppressor gene tropomyosin 1 (TPM1). *J. Biol. Chem.*, **282**, 14328-14336.

14. Baek,D., Villen,J., Shin,C., Camargo,F.D., Gygi,S.P. and Bartel,D.P. (2008) The impact of microRNAs on protein output. *Nature*, **455**, 64-71.

15. Easow,G., Teleman,A.A. and Cohen,S.M. (2007) Isolation of microRNA targets by miRNP immunopurification. *RNA.*, **13**, 1198-1204.

16. Zhang,L., Ding,L., Cheung,T.H., Dong,M.Q., Chen,J., Sewell,A.K., Liu,X., Yates,J.R., III and Han,M. (2007) Systematic identification of C. elegans miRISC proteins, miRNAs, and mRNA targets by their interactions with GW182 proteins AIN-1 and AIN-2. *Mol. Cell*, **28**, 598-613.

17. Beitzinger,M., Peters,L., Zhu,J.Y., Kremmer,E. and Meister,G. (2007) Identification of human microRNA targets from isolated argonaute protein complexes. *RNA. Biol.*, **4**, 76-84.

18. Karginov,F.V., Conaco,C., Xuan,Z., Schmidt,B.H., Parker,J.S., Mandel,G. and Hannon,G.J. (2007) A biochemical approach to identifying microRNA targets. *Proc. Natl. Acad. Sci. U. S. A*, **104**, 19291-19296.

19. Hendrickson,D.G., Hogan,D.J., Herschlag,D., Ferrell,J.E. and Brown,P.O. (2008) Systematic identification of mRNAs recruited to argonaute 2 by specific microRNAs and corresponding changes in transcript abundance. *PLoS. ONE.*, **3**, e2126.

20. Landthaler,M., Gaidatzis,D., Rothballer,A., Chen,P.Y., Soll,S.J., Dinic,L., Ojo,T., Hafner,M., Zavolan,M. and Tuschl,T. (2008) Molecular characterization of human Argonaute-containing ribonucleoprotein complexes and their bound target mRNAs. *RNA.*, **14**, 2580-2596.

21. Gibcus,J.H., Tan,L.P., Harms,G., Schakel,R.N., de Jong,D., Blokzijl,T., Moller,P., Poppema,S., Kroesen,B.J. and van den Berg,A. (2009) Hodgkin lymphoma cell lines are characterized by a specific miRNA expression profile. *Neoplasia.*, **11**, 167-176.

22. Ivanovska,I., Ball,A.S., Diaz,R.L., Magnus,J.F., Kibukawa,M., Schelter,J.M., Kobayashi,S.V., Lim,L., Burchard,J., Jackson,A.L. *et al.* (2008) MicroRNAs in the miR-106b family regulate p21/CDKN1A and promote cell cycle progression. *Mol. Cell Biol.*, **28**, 2167-2174.

23. Keene,J.D., Komisarow,J.M. and Friedersdorf,M.B. (2006) RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. *Nat. Protoc.*, **1**, 302-307.

24. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207-210.

25. Lee,I., Ajay,S.S., Yook,J.I., Kim,H.S., Hong,S.H., Kim,N.H., Dhanasekaran,S.M., Chinnaiyan,A. and Athey,B.D. (2009) New class of microRNA targets containing simultaneous 5'-UTR and 3'-UTR interaction sites. *Genome Res.*.

26. Hossain,A., Kuo,M.T. and Saunders,G.F. (2006) Mir-17-5p regulates breast cancer cell proliferation by inhibiting translation of AIB1 mRNA. *Mol. Cell Biol.*, **26**, 8191-8201.

27. Kiriakidou,M., Nelson,P.T., Kouranov,A., Fitziev,P., Bouyioukos,C., Mourelatos,Z. and Hatzigeorgiou,A. (2004) A combined computational-experimental approach predicts human microRNA targets. *Genes Dev.*, **18**, 1165-1178.

28. Sylvestre,Y., De Guire,V., Querido,E., Mukhopadhyay,U.K., Bourdeau,V., Major,F., Ferbeyre,G. and Chartrand,P. (2007) An E2F/miR-20a autoregulatory feedback loop. *J. Biol. Chem.*, **282**, 2135-2143.

29. Liu,Q., Fu,H., Sun,F., Zhang,H., Tie,Y., Zhu,J., Xing,R., Sun,Z. and Zheng,X. (2008) miR-16 family induces cell cycle arrest by regulating multiple cell cycle genes. *Nucleic Acids Res.*, **36**, 5391-5404.

30. Yeung,M.L., Yasunaga,J., Bennasser,Y., Dusetti,N., Harris,D., Ahmad,N., Matsuoka,M. and Jeang,K.T. (2008) Roles for microRNAs, miR-93 and miR-130b, and tumor protein 53-induced nuclear protein 1 tumor suppressor in cell growth dysregulation by human T-cell lymphotrophic virus 1. *Cancer Res.*, **68**, 8976-8985.

31. Volinia,S., Calin,G.A., Liu,C.G., Ambs,S., Cimmino,A., Petrocca,F., Visone,R., Iorio,M., Roldo,C., Ferracin,M. *et al.* (2006) A microRNA expression signature of human solid tumors defines cancer gene targets. *Proc. Natl. Acad. Sci. U. S. A*, **103**, 2257-2261.

32. Landais,S., Landry,S., Legault,P. and Rassart,E. (2007) Oncogenic potential of the miR-106-363 cluster and its implication in human T-cell leukemia. *Cancer Res.*, **67**, 5699-5707.

33. Johnson,S.M., Grosshans,H., Shingara,J., Byrom,M., Jarvis,R., Cheng,A., Labourier,E., Reinert,K.L., Brown,D. and Slack,F.J. (2005) RAS is regulated by the let-7 microRNA family. *Cell*, **120**, 635-647.

34. Tsang,W.P. and Kwok,T.T. (2008) Let-7a microRNA suppresses therapeutics-induced cancer cell death by targeting caspase-3. *Apoptosis.*, **13**, 1215-1222.

35. Bonci,D., Coppola,V., Musumeci,M., Addario,A., Giuffrida,R., Memeo,L., D'Urso,L., Pagliuca,A., Biffoni,M., Labbaye,C. *et al.* (2008) The miR-15a-miR-16-1 cluster controls prostate cancer by targeting multiple oncogenic activities. *Nat. Med.*, **14**, 1271-1277.

36. Mott,J.L., Kobayashi,S., Bronk,S.F. and Gores,G.J. (2007) mir-29 regulates Mcl-1 protein expression and apoptosis. *Oncogene*, **26**, 6133-6140.

37. Emmerich,F., Meiser,M., Hummel,M., Demel,G., Foss,H.D., Jundt,F., Mathas,S., Krappmann,D., Scheidereit,C., Stein,H. *et al.* (1999) Overexpression of I kappa B alpha without inhibition of NF-kappaB activity and mutations in the I kappa B alpha gene in Reed-Sternberg cells. *Blood*, **94**, 3129-3134.

38. Cabannes,E., Khan,G., Aillet,F., Jarrett,R.F. and Hay,R.T. (1999) Mutations in the IkBa gene in Hodgkin's disease suggest a tumour suppressor role for IkappaBalpha. *Oncogene*, **18**, 3063-3070.

39. Emmerich,F., Theurich,S., Hummel,M., Haeffker,A., Vry,M.S., Dohner,K., Bommert,K., Stein,H. and Dorken,B. (2003) Inactivating I kappa B epsilon mutations in Hodgkin/Reed-Sternberg cells. *J. Pathol.*, **201**, 413-420.

40. Weniger,M.A., Melzner,I., Menz,C.K., Wegener,S., Bucur,A.J., Dorsch,K., Mattfeldt,T., Barth,T.F. and Moller,P. (2006) Mutations of the tumor suppressor gene SOCS-1 in classical Hodgkin lymphoma are frequent and associated with nuclear phospho-STAT5 accumulation. *Oncogene*, **25**, 2679-2684.

41. Schmitz,R., Hansmann,M.L., Bohle,V., Martin-Subero,J.I., Hartmann,S., Mechtersheimer,G., Klapper,W., Vater,I., Giefing,M., Gesk,S. *et al.* (2009) TNFAIP3 (A20) is a tumor suppressor gene in Hodgkin lymphoma and primary mediastinal B cell lymphoma. *J. Exp. Med.*.

42. Maggio,E.M., van den Berg,A., de Jong,D., Diepstra,A. and Poppema,S. (2003) Low frequency of FAS mutations in Reed-Sternberg cells of Hodgkin's lymphoma. *Am. J. Pathol.*, **162**, 29-35.

43. Lal,A., Kim,H.H., Abdelmohsen,K., Kuwano,Y., Pullmann,R., Jr., Srikantan,S., Subrahmanyam,R., Martindale,J.L., Yang,X., Ahmed,F. *et al.* (2008) p16(INK4a) translation suppressed by miR-24. *PLoS. ONE.*, **3**, e1864.

44. Licatalosi,D.D., Mele,A., Fak,J.J., Ule,J., Kayikci,M., Chi,S.W., Clark,T.A., Schweitzer,A.C., Blume,J.E., Wang,X. *et al.* (2008) HITS-CLIP yields

genome-wide insights into brain alternative RNA processing. *Nature*, **456**, 464-469.

45. Chi,S.W., Zang,J.B., Mele,A. and Darnell,R.B. (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*.
46. Rigoutsos,I. (2009) New Tricks for Animal MicroRNAs: Targeting of Amino Acid Coding Regions at Conserved and Nonconserved Sites. *Cancer Res.*.
47. Hammell,M., Long,D., Zhang,L., Lee,A., Carmack,C.S., Han,M., Ding,Y. and Ambros,V. (2008) mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts. *Nat. Methods*, **5**, 813-819.

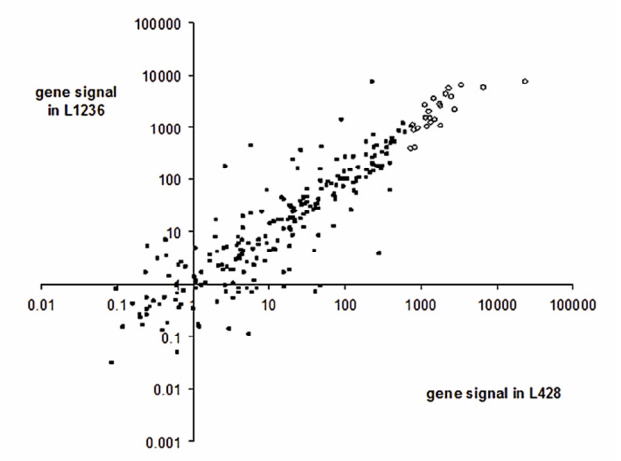
Supplementary data 1:

The Excel file is available for download at Oxford Journals using the following address: *http://nar.oxfordjournals.org/content/suppl/2009/09/01/gkp715.DC1*

Supplementary data 2:

Top 5% miRNAs in HL cell lines

Top 5% miRNAs	**Seed matching site**	
miRNA family	**8mer**	**6mer**
let-7afi/98	ctacctca	tacctc
miR-15b/16/195	tgctgcta	gctgct
miR-17/20a/93/106ab	gcacttta	cacttt
miR-21	ataagcta	taagct
miR-25/92	gtgcaata	tgcaat
miR-19ab	tttgcaca	ttgcac
miR-29ab	tggtgcta	ggtgct
miR-103/107	atgctgca	tgctgc
miR-142-5p	actttata	ctttat
miR-142-3p	acactaca	cactac



Expression levels of the top 5% (23/470) miRNAs in two Hodgkin lymphoma cell lines, L1236 and L428 according to Gibcus et al. 2009. Open circle, top 5% expressed miRNAs. Black square, the remaining 95%.

Supplementary data 3: Cloning for luciferase reporter assays

Gene name	ACCESSION. Version	Forward primer
ADRB2	NM_000024.4	5' GAGCTC-CCCCCAACAGAACACTAAAC 3'
CCL1	NM_002981.1	5' GAGCTC- GCAGATTTCTTTCCATTGTG 3'
CD274	NM_014143.2	5' GAGCTC-CTTCTGATCTTCAAGCAGGG 3'
FBXO31	NM_024735.2	5' GAGCTC-GAACTCTGACCTGTGAATAG 3'
GPR137B	NM_003272.2	5' GAGCTC-AAGCCTTGGGTAGCATCAG 3'
NPAT	NM_002519.2	5' GAGCTC-GTGTAGGGAATGGGATATTGAC 3'
OBFC2A	NM_001031716.1	5' GAGCTC-CATGCCCTACTTGAACAC 3'
RAB12	NM_001025300.2	5' GAGCTC-GTCCGATGCTGTTGATTTCC 3'
RBJ	NM_016544.1	5' GAGCTC-CCAGAGCGTTGCTTTATC 3'
SGK1	NM_005627.3	5' GAGCTC-CAGCTGACAGGACATCTTAC 3'
YES1	NM_005433.3	5' CTCGAG-GGTAAACTGGAATCCCAGATATGG 3'
ZNF22	NM_006963.3	5' GAGCTC- AGTCCAGCTACCCTCATTTC 3'
ZNFX1	NM_021035.2	5' GAGCTC-GCTCCCTAATGAAGGAACTG 3'

Gene name	ACCESSION. Version	Reverse primer
ADRB2	NM_000024.4	5’ GCGGCCGC- AACAACTGAAGCTGCTCCTC 3'
CCL1	NM_002981.1	5’ GCGGCCGC-GTTGGGGGTTGATGATTGTA 3'
CD274	NM_014143.2	5’ GCGGCCGC- TCCATGTTTTACTAGATGTGAG 3'
FBXO31	NM_024735.2	5' GCGGCCGC-CGCAGGTGTTAACACAACATGG 3'
GPR137B	NM_003272.2	5' GCGGCCGC- CTGAGGGCTCATTAGAGTC 3'
NPAT	NM_002519.2	5' GCGGCCGC-AAGGTTCAGCAATTAAGAAGAC 3'
OBFC2A	NM_001031716.1	5' GCGGCCGC-TGACTCACCTCCAGTATG 3'
RAB12	NM_001025300.2	5' GCGGCCGC-GGCCATGAATGGAGCTTTG 3'
RBJ	NM_016544.1	5' GCGGCCGC-TCTCTTGGGTAGCAACAC 3'
SGK1	NM_005627.3	5' GCGGCCGC-ATGGGATGAGGGAAGGATTG 3'
YES1	NM_005433.3	5' GCGGCCGC-TAGGTGCATTCAATGAGAAC 3'
ZNF22	NM_006963.3	5' GCGGCCGC- CTGAGGGCTCATTAGAGTC 3'
ZNFX1	NM_021035.2	5' GCGGCCGC-ATTGGCTTTATAAGCTAAAGTG 3'

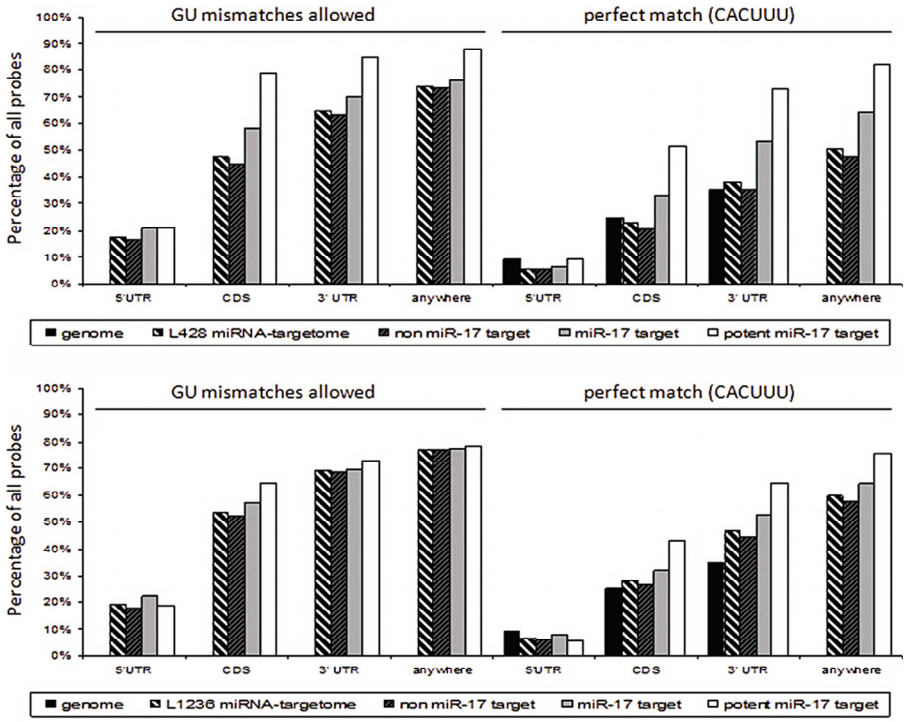
Restriction site	sequence
SacI	GAGCTC
XhoI	CTCGAG
NotI	GCGGCCGC

Supplementary data 4: Increasing percentage of probes with seed matching sites for miR-17 seen in "depleted" probe sets of cells transfected with anti-miR-17/20/93/106

			8mer site		6mer site	
	Depletion fold	**# probes**	**miR-220**	**miR-17**	**miR-220**	**miR-17**
genome	-	49516	0.4%	3%	25%	35%
L428	-	2703	0.4%	6%	21%	38%
+ anti-miR-220	not depleted	2486	0.6%	6%	21%	39%
	≥ 2	217	0%	7%	28%	51%
	≥ 3	37	0%	8%	22%	41%
+ anti-miR-17/20/93/106	not depleted	2210	0.4%	5%	21%	35%
	≥ 2	493	0.2%	14%	24%	53%
	≥ 3	100	0%	28%	26%	58%
	≥ 4	33	0%	39%	39%	73%
L1236	-	3164	0.5%	10%	27%	47%
+ anti-miR-17/20/93/106	not depleted	2269	0.4%	8%	26%	45%
	≥ 2	895	0.8%	14%	30%	53%
	≥ 3	399	0.8%	15%	31%	53%
	≥ 4	206	1.0%	12%	34%	56%
	≥ 5	114	1.8%	15%	39%	61%
	≥ 6	62	0%	21%	34%	66%
	≥ 7	37	0%	27%	30%	65%

Note that the actual probes with seed matching sites are labeled in Supplementary 1.

Supplementary data 5



In almost all analysis, "potent miR-17 target" always has the highest percentage of all probes with miR-17 seed matching sites. Perfect seed matching site in 3' UTR appeared to be the best discriminator for miR-17 targets from the genome. Note that the actual probes with seed matching sites are labeled in Supplementary data 1.

Chapter 6

Summarizing discussion

Introduction

During the last decade, RNA interference (RNAi) has become an important tool to investigate the function of genes and the gene-gene interaction between them. By adding double stranded RNA (dsRNA) constructs, the RNAi pathway induces downregulation of the expression of a gene [1]. This is done by administering dsRNAs with sequence homology against the target gene. The dsRNAs inflict a cascade of reactions ending in the destruction of the messenger RNA (mRNA) originating from the target gene, and as a consequence the protein is not formed. By studying the phenotype induced by downregulation of the mRNA, the function of the chosen gene is elucidated. Unfortunately, RNAi may also have collateral effects, of which the so called off-targets are the most important. Off-target effects can occur because of the presence of sequence similarity between the target genes and other, genes.

RNAiSelect; a novel bioinformatics tool that profiles off-targets

With time, knowledge grew about how RNAi works and what the limitations are in using them for experiments [2]. One of these limitations is that off-targets may occur when an RNAi molecule has similarity to other regions on the genome than its intended target [3]. Bioinformatics plays a key role in circumventing these off-targets by analysing the sequence of the RNAi molecule and reporting about the uniqueness amongst genes [4,5,6]. Still, one important property of RNAi was ignored by the current available tools: The RNAi machinery can not only target mature RNA sequences but can also target intron containing pre-messenger RNA and other nuclear located RNAs [7,8,9]. This has major consequences for the generality of the already existing bioinformatics tools, as these are only considering mature messenger RNA (mRNA) present in the cytoplasm. The knowledge that the RNAi machinery can also have a nuclear localization, indicates that there are many more sequences that can be the target of regulatory breakdown.

Our tool “RNAiSelect” (chapter 2) responds to these observations by allowing a much more thorough analysis including the complete genome to find off-targets. RNAiSelect is based on a novel algorithm that requires only seconds to complete a comprehensive search of 21-nt against the complete genome while allowing up to 3 mismatches. By analysing the complete genome, pre-mRNA sequences are also considered as well as other known

(i.e. miRNAs) or unknown regions of the genome. This approach certainly has ample false-positive hits caused by sequences that do have similarity to the RNAi molecule under investigation, but do not inflict a real biological effect. **However, it is important to realize that the purpose of RNAiSelect is not to identify off-targets *per se*, but to select the best RNAi molecule candidates amongst the many that are possible.** By analysing all candidate siRNAs that may be created for a particular target gene, the ones with the least number of predicted potential off-targets will be the most preferable ones and from which the most suitable according to efficiency properties may be chosen. In addition, the number of mismatches against predicted potential off-targets as well as the type of mismatch is considered resulting in a general score that also aids in choosing the best siRNA with the least number of potential off-targets.

We have validated our approach using publically available micro-arrays. By analysing the off-target profiles of the used dsRNAs and after comparing the output with the actual regulatory differences on the arrays, we concluded that RNAiSelect indeed has a predictive value regarding which sequences can contribute to off-targets. Although the array showed numerous marginal expression differences considering the complete array, a significant effect was observed for the set of predicted off-targets. The knowledge that small changes at the mRNA level can have large effects on protein expression levels underscores the significant effect we have found in chapter 2 [10].

Fortunately, in our genome-wide analysis of all potential siRNA sequences using RNAiSelect, we have found that the majority of genes have potent siRNAs with relatively few predicted potential off-targets (up to 80% less than average). When performing a genome-wide screen using dsRNA molecules, RNAiSelect will be particularly useful to identify these clean RNAis and assists in the decision which of them are most “clean” and therefore most promising to proceed with. When designing a RNAi experiment, one can choose a single siRNA (21 nt in length) instead of a large dsRNA (approximately between 250 and 800 nt in length) to prevent as many potential off-targets as possible. In addition, the outcome (especially micro-array data) of already performed RNAi experiments can now be re-evaluated using RNAiSelect and it can be checked whether reported RNAi-induced effects are *bona fide* effects or can be explained by off-target effects.

A solution to the overlooked problem of off-targets shared by independent dsRNAs targeting the same gene

It is currently not possible to predict with a high level of accuracy whether a specific RNAi molecule does provoke a potent on-target effect and it is also not possible to predict whether an off-target effect, identified by bioinformatic tools will indeed provoke a measurable biological effect. A method to circumvent the usage of bioinformatics tools is to use two independent and non-overlapping siRNA constructs to downregulate the same gene of interest. The usage of this type of controlled-experiment is reasonable because the assumption is that two completely different dsRNAs will have identical on-target effects but will have a completely different genome wide off-target profile. Any identical biological effect provoked by the distinct dsRNA constructs is considered as a *bona fide* on target effect.

Unfortunately, this assumption has proven to be very wrong (chapter 3). Statistical analysis and our results show that within the *Drosophila* genome it is highly likely for distinct 21 nt sequences derived from the same gene to have closely mapped sequence similarity elsewhere on the genome. In other words, if one slices a Drosophila gene in pieces of 21-nt and a sequence similarity analysis is performed, many of the individual 21-nt pieces will find homology within the same 'other' *Drosophila* gene (i.e. potential off-target gene). **Indeed, our results using actual genome data show that most genes have many of these sequences that map to the same potential off-target. We therefore concluded that an identical biological effect induced by randomly choosing distinct dsRNAs derived from one gene is not per definition a *bona fide* on-target effect.** Moreover, our results show that even when 3 or more distinct dsRNAs are used, there is still a significant possibility of overlapping off-targets which means that "blindly" picking dsRNA constructs is almost never suffice to counter off-target effects.

Despite this knowledge, the use of a double controlled RNAi experiment is appealing as it may indeed prevent many false observations. RNAiSelect (presented in chapter 3) is useful in finding the best predicted siRNA candidate combinations that have most likely no shared off-targets based on sequence similarity. In summary, we present a method to identify 2 distinct dsRNAs from a gene of choice that do not show any potential off-target overlap, -based on sequence similarity- by performing a thorough potential

off-target overlap analysis. This tool is freely available at http://www.rnaiselect.info/dsrna and may be used by the *Drosophila* community where dsRNAs are generally used for gene down-regulation studies.

Pantethine rescues dPANK depleted Drosophila S2 cells involved in CoA metabolism

The procedure of using RNAi in researching a human disease model in *Drosophila* has proven to be very useful to unravel the mechanisms behind the disease [11]. RNAiSelect has been used in chapter 4 to generate clean and specific dsRNAs to model the human disease PKAN (Pantothenate kinase–associated neurodegeneration) in *Drosophila*. PKAN is a rare neuronal disease caused by a mutation in the human PANK2 gene.

PANK2 is evolutionary conserved amongst many species and is essential in the biochemical pathway that converts vitamin B5 into Coenzyme A (CoA) [12]. *Drosophila* has proven to be particularly suitable as a model for PKAN, as it shares many characteristics that are reported in human PKAN patients carrying a mutation in PANK2. These phenotypes comprise increased loss of locomotor function, neurodegeneration, and a decreased lifespan.

With RNAi, a cell based PKAN *Drosophila* model (S2) was created. A biochemically measurable effect of this knock-down mutant cell line is the severely decreased levels of CoA. In addition, a phenotypical effect is measured through decreased cell-count in time. By adding the chemical substrate pantethine to the growth medium, CoA levels and cell numbers were restored comparable to wild type S2 cells. Thus our research suggests that an alternative route may exist leading from pantethine to CoA, independent from pantothenate kinase. This hypothesis is further supported by our findings that any residual *dPank/fbl* kinase activity does not on itself contribute to the phosphotransferase conversion of Pantethine, because (1) western blotting demonstrated the nearly complete knock-down of dPank/fbl and (2) the addition of increasing concentrations of vitamin B5 did not have any significant effect on the mutant phenotype.

Pantethine has also been administered to *Drosophila* fly mutants, which rescues all tested aspects of their neurodegenerative phenotype and increased their life span. Moreover, it was demonstrated that in a human cell

model for PKAN, pantethine also works protective. Together our results show that a well-controlled RNAi experiment in Drosophila S2 cells leads to the identification of a lead compound, that may be at the base of a possible treatment for a so far non-treatable disease.

Identifying miRNA targets in human cells to understand Hodgkin Lymphoma

In addition to the possibility of our algorithm to search for full length sequence similarities, short 6-nt 'seed' sequences that can result in possible miRNA-like effects can be searched for. The latter is also useful for identifying or confirming naturally occurring miRNA targets that regulate endogenous genes. This is particularly interesting for research concerning the role of miRNAs in developmental regulation, but also to study the effect of malfunctioning miRNAs as they occur in several cancerous diseases like Hodgkin Lymphoma (HL; a cancer originating from leucocytes).

Chapter 5 describes a high-throughput approach to identify endogenous miRNA targets of untreated human cells in which specific miRNAs are upregulated. The approach uses a combination of wet experiments (in vitro/vivo studies) and RNAiSelect (in silico) capabilities to find sequence similarities. While other methods have been established to find miRNA targets, none of them can be done in a wild type background because these published methods make use of cells that overexpress synthetic miRNAs and flag-tagged miRNA related proteins which require a significant modulation of the cells [13,14,15,16,17]. These approaches clearly influence the physiological state of the cells under investigation and will affect the outcome of the experiments. This underscores the need for a better system. Our high throughput approach (as described in chapter 5) is not altering physiological conditions and is accomplished by using antibodies against Argonaute (Ago2; which is the catalytic component of the RISC complex that binds and cleaves targeted mRNAs by complementary binding with the miRNA seed region) in combination with immuno-precipitation of this RISC complex. The precipitated complex contains mRNAs that were bound to RISC and are therefore candidate miRNA targets. Using this method, miRNA targets bound to Ago2 can be isolated from untreated and physiologically relevant cell or tissue samples and identified by nucleotide sequencing.

A proof of principle has been performed by using 2 different HL cell lines that have several characteristic miRNAs upregulated (miR-17/20/93/106). First, RISC from untreated HL cells was isolated and attached mRNAs were isolated, sequenced and identified. Next, the HL cell lines have been transfected with anti-miRNAs (i.e. anti-miR-17) and again pools of mRNAs that were attached to Ago2 have been isolated. By cross-referencing both pools, mRNAs that are bound to RISC in the unaltered cells but not in transfected cells could be isolated and were candidates for being actual targets of the miRNAs under investigation.

Although this pool is large and contains many false positives due to non-specific or non-related binding of mRNAs to RISC, it is a perfect start for RNAiSelect to condense the list of candidate genes to a validated list based on seed complementarity to the miRNAs under inverstigation. By using RNAiSelect, a list of mRNAs containing highly significant enrichment of seed-sequences against the miRNAs under investigation has been isolated from the large pool of mRNA that were found in the Ago2 immune-precipitate. Because these mRNAs are both found in RISC and have multiple occurrences of the seed sequences against miRNAs that are known to be highly active, it is very likely for them to be actual targets. Indeed, by using luciferase assays against 8 chosen genes from the isolated mRNA set predicted to be downregulated by miRNA-17, all of them are confirmed to be upregulated upon miRNA-17 inhibition. Notably, 2 genes have also been assayed using the same method, which showed significant enrichment in the Ago complex, but appeared to lack the seed region of miRNA-17 after analyzing with RNAiSelect. Upon miRNA-17 inhibition, the correlation with this miRNA could not be confirmed consisting with the lacking seed region. This observation demonstrates that initial large pools of mRNAs could effectively be reduced with RNAiSelect by reducing false positives and therefore provides the necessary sensitivity of the novel high-throughput approach to identify miRNA targets.

This thesis has also given new insights into how miRNA-mRNA interactions occur. Studies from the Bartel group have shown that 8-mer seed sequences show a higher and more reliable occurrence of interaction, suggesting a preference for these 8-mer interactions [18]. Our results however show that although 8-mer sites in 3'-UTR regions definitely are good indicators to identify miRNA-specific targets, they are not obligatory for effective targeting. In reality, 5 out of 9 targets of the miR-17 seed

family were confirmed with the luciferase assay, but did not contain 8-mer sites as opposed to the 6-mers.

Summarizing conclusion

Gene expression regulation by RNAs is part of a complex machinery for which recently most has been unraveled. Roughly this topic can be divided in endogenous RNA regulation and exogenous RNAi research and examples of both of them have been under investigation and discussed in detail in this thesis. First bioinformatics tools were designed to be able to evaluate results obtained with manipulating regulation by RNA (chapter 2 and 3) and subsequently these tools were implemented in 'wet' experiments (chapter 4 and 5). For exogenous siRNAs or derivatives thereof, novel methods have been devised that aid in designing clean RNAi experiments to identify a lead compound suitable to develop a future therapy for a so far non-curable disease. For the endogenous regulation, a novel method has been presented to deduce natural miRNA targets in order to interpret and further direct 'wet's experiments highly efficiently. The results presented in this thesis underscore the importance of merging bioinformatics in medical relevant life science research..

References

1. Misquitta L, Paterson BM (1999) Targeted disruption of gene function in Drosophila by RNA interference (RNA-i): A role for nautilus in embryonic somatic muscle formation. Proceedings of the National Academy of Sciences of the United States of America 96: 1451-1456.
2. Maine EM (2001) RNAi As a Tool for Understanding Germline Development in Caenorhabditis elegans: Uses and Cautions. Developmental Biology 239: 177-189.
3. Jackson A, Bartz S, Schelter J, Kobayashi S, Burchard J, et al. (2003) Expression profiling reveals off-target gene regulation by RNAi. Nature biotechnology 21: 635-637.
4. Yamada T, Morishita S (2005) Accelerated off-target search algorithm for siRNA. Bioinformatics 21: 1316.

5. Qiu S, Adema C, Lane T (2005) A computational study of off-target effects of RNA interference. Nucleic Acids Res 33: 1834-1847.
6. Naito Y, Yamada T, Matsumiya T, Ui-Tei K, Saigo K, et al. (2005) dsCheck: highly sensitive off-target search software for double-stranded RNA-mediated RNA interference. Nucleic Acids Res 33: W589-591.
7. Lin S-L, Kim H, Ying S-Y (2008) Intron-mediated RNA interference and microRNA (miRNA). Frontiers in bioscience : a journal and virtual library 13: 2216-2230.
8. Liu J, He Y, Amasino R, Chen X (2004) siRNAs targeting an intronic transposon in the regulation of natural flowering behavior in Arabidopsis. Genes & development 18: 2873-2878.
9. Bosher J, Dufourcq P, Sookhareea S, Labouesse M (1999) RNA Interference Can Target Pre-mRNA: Consequences for Gene Expression in a Caenorhabditis elegans Operon. Genetics 153: 1245-1256.
10. Gygi SP, Rochon Y, Franza BR, Aebersold R (1999) Correlation between Protein and mRNA Abundance in Yeast. Molecular and Cellular Biology 19: 1720-1730.
11. Botas J (2007) Drosophila researchers focus on human disease. Nat Genet 39: 589-589.
12. Zhou B, Westaway SK, Levinson B, Johnson MA, Gitschier J, et al. (2001) A novel pantothenate kinase gene (PANK2) is defective in Hallervorden-Spatz syndrome. Nat Genet 28: 345-349.
13. Easow G, Teleman AA, Cohen SM (2007) Isolation of microRNA targets by miRNP immunopurification. RNA 13: 1198-1204.
14. Zhang L, Ding L, Cheung TH, Dong M-Q, Chen J, et al. (2007) Systematic Identification of C. elegans miRISC Proteins, miRNAs, and mRNA Targets by Their Interactions with GW182 Proteins AIN-1 and AIN-2. Molecular cell 28: 598-613.
15. Karginov FV, Conaco C, Xuan Z, Schmidt BH, Parker JS, et al. (2007) A biochemical approach to identifying microRNA targets. Proceedings of the National Academy of Sciences 104: 19291-19296.
16. Hendrickson DG, Hogan DJ, Herschlag D, Ferrell JE, Brown PO (2008) Systematic Identification of mRNAs Recruited to Argonaute 2 by Specific microRNAs and Corresponding Changes in Transcript Abundance. PLoS ONE 3: e2126.
17. Landthaler M, Gaidatzis D, Rothballer A, Chen PY, Soll SJ, et al. (2008) Molecular characterization of human Argonaute-containing ribonucleoprotein complexes and their bound target mRNAs. RNA 14: 2580-2596.
18. Grimson A, Farh KK-H, Johnston WK, Garrett-Engele P, Lim LP, et al. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Molecular cell 27: 91-105.

Chapter 7

Nederlandse Samenvatting

Nederlandse samenvatting

Een belangrijk onderdeel binnen de (bio)wetenschap is onderzoek naar de functie van genen. Voorheen werd dat vooral gedaan door opzettelijk mutaties in genen te maken, waardoor uiteindelijk specifieke eiwitten fouten bevatten en niet goed werken. De consequenties van disfunctionerende eiwitten kunnen daarna bestudeerd worden in cellen of in modelorganismen. De geconstateerde gebreken correleren meestal goed met de aangetaste functie van het gemuteerde gen. Logisch gevolg is dat er sterke aanwijzingen ontstaan over de normale werking en functie van het eiwit die door het gen in kwestie wordt gecodeerd. Tegenwoordig wordt er echter veel vaker een nieuwere techniek gebruikt: RNA interference, of kortweg RNAi. Met de RNAi techniek wordt niet het DNA gemuteerd, maar wordt het RNA afgebroken. Omdat deze techniek centraal staat binnen dit proefschrift, zal ze hier eerst uitgelegd worden.

RNAi: een revolutionaire technologie

In vrijwel alle organismen bestaat een geconserveerde route die globaal bestaat uit 4 niveaus. Het begint met de informatie die ligt opgeslagen in het DNA van de celkern. Deze informatie is omsloten in genen en bevat de bouwinstructies om eiwitten te maken. De in alle cellen aanwezige ribosomen lezen deze instructies af en maken het uiteindelijke eiwit. Dat doen ze echter niet direct vanaf het origineel (DNA), maar via de tussenvormen die **pre-m**RNA en **m**RNA worden genoemd. Het pre-mRNA is een exacte kopie van het gen waar nog laatste wijzigingen op kunnen worden gemaakt voordat het eiwit wordt gemaakt. Het gewijzigde RNA is het messenger RNA, of mRNA. De 4 niveaus kunnen samengevat worden in het volgende schema:

(1) DNA -> (2) -> pre-mRNA -> (3) mRNA -> (4) Eiwit.

In het verleden was vaak het eerste niveau ‘DNA’ het doelwit om een eiwit uit te schakelen, maar tegenwoordig wordt de RNAi techniek gebruikt. RNAi grijpt aan op de tussenvorm op de weg van genen naar eiwitten, namelijk het RNA. Waar voorheen genen werden gemuteerd op niveau (1), wordt met RNAi 'slechts' geïnterfereerd op niveau (3), zodat het eiwit op niveau (4) nooit wordt gemaakt. Deze manier van interfereren kan

theoretisch hetzelfde effect hebben als het uitschakelen van de aanmaak van het eiwit op niveau 1. Zonder handleiding kan er immers nooit een eiwit gebouwd worden. Het grootste voordeel van de RNAi techniek is dat deze heel gericht toegepast kan worden binnen cellen, maar juist ook in een volwassen modelorganisme. Dit in tegenstelling tot de vroegere technieken die mutaties op DNA niveau introduceerden. Om mutaties op DNA niveau te introduceren zijn meer stappen nodig en soms ook meerdere generaties van modelorganismen, terwijl RNAi op elk gewenst moment en heel gericht kan worden toegepast. Niet voor niets vindt RNAi daarom niet alleen steeds meer zijn weg in (fundamenteel) wetenschappelijk onderzoek, maar inmiddels ook in klinische toepassingen om bijvoorbeeld genen uit te schakelen die te maken hebben met de groei van kankergezwellen. De invloed van de RNAi techniek is dermate groot gebleken dat de ontdekkers van deze techniek, Fire en Mello, inmiddels de Nobelprijs hebben gekregen voor hun werk.

RNAi: hoe werkt het en wat zijn de nadelen?

Het mechanisme achter RNAi is relatief eenvoudig en begint met het binnenbrengen van een (synthetisch) dubbelstrengs RNAi molecuul in de cel. Deze dissocieert vervolgens in 2 enkelstrengs RNAs middels een actief proces. Een enkelstrengs RNA kan gemakkelijk met complementaire mRNA moleculen binden (niveau 3 van de vorige paragraaf) zodat het weer een dubbelstrengs RNA gaat vormen. Als dit gebeurt, dan wordt dit herkend door de cel als iets abnormaals. mRNA is namelijk enkelstrengs zodat het zonder belemmering afgelezen kan worden door de ribosomen (de eiwitfabriekjes die het mRNA als handleiding gebruiken om niveau 4 te bereiken). Een mRNA die nu (deels) dubbelstrengs is geworden, wordt actief afgebroken met als gevolg dat het eiwit nooit gevormd zal kunnen worden.

DNA en RNA moleculen bestaan uit bouwstenen, ook wel de ‘basen’ genoemd. Iedere succesvolle RNAi poging wordt uiteindelijk uitgevoerd met een sequentie met een lengte van +/- 21 basen. De oorspronkelijk aangeboden lengte van het RNAi molecuul maakt niet uit, als de fruitvlieg (*Drosophila melanogaster*) wordt gebruikt als modelorganisme. In dat geval worden meestal relatief lange RNAi moleculen gebruikt, de zogenaamde “dsRNAs” van +/- 350 basen lang. Een in de natuur voorkomend eiwit, genaamd DICER zal deze uiteindelijk alsnog gaan knippen in de kortere +/-

21-bp variant (de siRNAs, ofwel short interfering RNA) en al deze siRNAs kunnen actief bijdragen aan de RNAi reactie. Daar komt dan ook direct een probleem om de hoek kijken: de basen van DNA/RNA omvatten maar 4 varianten, dus hoe specifiek kan een siRNA van 'slechts' 21 basen zijn ten opzichte van het gehele genoom (de set van genen in de kern van een cel), dat in het geval van de fruitvlieg uit maar liefst 168 miljoen basen bestaat? Met andere woorden, hoe kunnen we er zeker van zijn dat de relatief korte sequentie van +/- 21 basen daadwerkelijk alleen op het bedoelde gen gaat aangrijpen en niet ook op een ander gen dat toevallig dezelfde sequentie heeft? Als men het aantal mogelijke combinaties uitrekent ($4^{\wedge 21}$), dan blijkt dat aantal heel groot en is de kans klein dat een sequentie van +/- 21 bp van de siRNA toevallig ergens anders in het genoom ook voorkomt. Er is echter uit studies gebleken dat de 21 basen niet exact overeen hoeven te komen, omdat de cel het vaak toestaat dat er foutjes, de zogenaamde 'mismatches', in voorkomen. Juist door de tolerantie van deze mismatches wordt de specificiteit een heel stuk kleiner en is het niet ondenkbaar dat een enkel stuk siRNA op meerdere plekken van het genoom een hoge mate van similariteit kan vinden. Inderdaad blijkt in de praktijk dat er meerdere genen onbedoeld aangedaan kunnen zijn door een RNAi behandeling. Dit effect wordt ook wel het off-target effect genoemd. Dit fenomeen kan vervelende gevolgen hebben voor de wetenschap, want dit kan er voor zorgen dat er verkeerde conclusies getrokken worden. Maar een off-target effect kan ook ernstige gevolgen hebben als RNAi gebruikt wordt bij een klinische toepassing. In dit geval kan niet alleen het zieke gen uitgeschakeld worden maar ook nog ongewild andere genen, waardoor de situatie van de patiënt theoretisch juist kan verslechteren. Gezien het bovenstaande is het dus essentieel om off-target effecten te kunnen minimaliseren.

Voorkómen van de nadelen

De bioinformatica discipline heeft geprobeerd om software te maken die de kans op off-targets vermindert door meer gericht siRNAs te ontwerpen. De sequentie van een siRNA wordt dan met behulp van een computer vergeleken met alle voorkomende sequenties van de mRNAs (niveau 3) zoals deze door genen worden gemaakt. Met de uitkomst van die analyse kan bij voorbaat al redelijk ingeschat worden of een bepaalde siRNA specifiek zal zijn of wellicht off-targets kan gaan veroorzaken. Doordat dergelijke analyses voor organismen die veel genen bevatten behoorlijk veel computerkracht vergen, zijn geavanceerde algoritmes gemaakt om de

analyse te versnellen. Op deze manier kan de onderzoeker binnen redelijke tijd een antwoord krijgen op de vraag of een bepaalde siRNA specifiek is of juist niet. Op basis van de eerste bevindingen van de werking van RNAi, werd ook al snel besloten om grote delen van het genoom uit te sluiten voor dit soort analyses omdat die niet vatbaar zou zijn voor de techniek. Het gaat daarbij dan voornamelijk om DNA sequenties die tussen de genen inliggen en niet coderen voor eiwitten. Door het daadwerkelijk aantal sequenties te beperken, die daarom moesten worden geanalyseerd, kon het gehele berekeningsproces flink worden versneld.

Verdere optimalisaties konden worden verkregen door te proberen nog meer sequenties uit te sluiten. Om die reden werd gekeken of specifieke gebieden binnen de genen zelf onaantastbaar zouden zijn voor RNAi en wellicht ook uitgesloten konden worden voor analyse. Een gen op het DNA binnen een celkern bestaat uit 3 type gebieden: 1) UTR (dit gedeelte van het gen zit aan de uiteinden en wordt niet vertaald naar mRNA), 2) exonen (dit is het deel van het gen dat daadwerkelijk wordt vertaald naar het mRNA) en 3) de intronen (dit is het deel van het gen dat niet bijdraagt aan de sequentie van het corresponderende mRNA). Deze intronen worden tijdens de vertaling naar het RNA als het ware weggegooid omdat ze uit de premature versie van het RNA (het pre-mRNA worden weggeknipt (dit gebeurt tijdens de hierboven genoemde overgang van niveau 2 naar 3). Vroeger noemde men deze intronen dan ook wel het 'junk-DNA', hoewel inmiddels bekend is dat deze intronen wel degelijk een functie hebben, maar daarover later meer. Er wordt in het algemeen aangenomen dat RNAi niet op niveau 2 (pre-mRNA), maar pas op niveau 3 (mRNA) aangrijpt. Deze aanname is gebaseerd op de lokalisatie van de hele RNAi machinerie en zal hieronder toegelicht worden. In een cel ligt het DNA opgeslagen binnen een compartiment (de celkern). Alles buiten de celkern wordt het cytoplasma genoemd; hierin zitten alle andere celorganellen (zoals de eerder genoemde 'eiwitfabriekjes' ribosomen). De vertaling van DNA en rijping naar mRNA vindt plaats binnen de celkern en daarna wordt het mRNA geëxporteerd naar het cytoplasma. In het cytoplasma wordt het mRNA afgelezen door de ribosomen en vervolgens ook het eiwit gemaakt. Onderzoek suggereerde dat de componenten van RNAi alleen in het cytoplasma aanwezig waren en in het cytoplasma wordt het mRNA afgebroken voordat het kan worden vertaald naar eiwitten via de ribosomen. Gebaseerd op de uitkomsten van dit onderzoek werd de conclusie getrokken dat RNAi alleen op niveau 3 kan aangrijpen (mRNA). Bioinformatici hebben gebruik gemaakt van deze

observatie. Zij hebben in hun analyses alleen de UTR en exonsequenties van genen meegenomen en al het overige uitgesloten. Door deze filtering hoeft er veel minder berekend te worden waardoor de analyses om potentiële off-targets in kaart te brengen veel sneller uitgevoerd konden worden. Door het digitaal inkorten van het genoom wordt dus een flinke tijdswinst behaald op de analyses. De winst maakt het mogelijk om met een hoge snelheid sequenties te kunnen analyseren zodat deze computer programma's ook geschikt zijn voor online toepassingen op het internet.

Zijn de aannames correct?

Inmiddels zijn er steeds meer aanwijzingen dat de hele RNAi machinerie niet alleen in het cytoplasma, maar ook in de celkern aanwezig is. Dat kan drastische consequenties hebben voor bestaande analyses met betrekking tot de specificiteit van RNAi, aangezien tot op heden vooral alleen naar de mRNA sequenties is gekeken en niet naar de volledige pre-mRNA sequenties. Daarnaast zijn er mechanismen blootgelegd die verwant zijn aan RNAi, maar het DNA in de kern als doelwit hebben in plaats van het RNA. Een analyse is daarom pas werkelijk compleet als er rekening wordt gehouden met alle mogelijke sequenties waarop een off-target zich kan voordoen. Hoofdstuk 2 presenteert een nieuwe bioinformatica methode die het toch mogelijk maakt om genoom-breed te zoeken naar off-targets inclusief eventuele mismatches. Daarbij wordt gebruik gemaakt van een nieuw programma, genaamd RNAi-Select, om dit toch heel snel te kunnen uitvoeren en de benodigde tijd voor een enkele analyse daarmee te beperken tot iets meer dan een seconde. Door deze snelheid zijn on-line toepassingen en/of bulkanalyses op vele siRNAs of dsRNAs mogelijk en deze komen ook uitgebreid aan de orde in dit hoofdstuk.

Door gebruik te maken van deze nieuwe methode om off-targets op te sporen zijn we erachter gekomen dat het aantal potentiële off-targets voor sommige RNAi moleculen soms verrassend hoog is. Na dit vastgesteld te hebben rees de vraag of deze hoge aantallen potentiële off-targets ook consequenties kunnen hebben voor de technieken en controles die vandaag de dag gebruikt worden binnen het RNAi onderzoek. Hoofdstuk 3 gaat hier specifiek op in en onderzoekt een methode die onderzoekers momenteel gebruiken om off-targets te minimaliseren in het modelorganisme Drosophila melanogaster (fruitvlieg). Om aan te tonen dat een bepaald RNAi construct specifiek is, gebruiken deze onderzoekers twee

onafhankelijke unieke RNAi constructen die beide bedoelt zijn om hetzelfde gen uit te schakelen. De eenvoudige gedachtegang daarachter is dat als twee totaal verschillende RNAi's onafhankelijk en in aparte experimenten van elkaar hetzelfde gen uitschakelen en ook dezelfde resultaten vertonen in het onderzoek, dat dan deze resultaten louter en alleen veroorzaakt worden door het uitschakelen van het doelgen. Dit lijkt een redelijke veronderstelling, want intuïtief wordt er van uitgegaan dat de unieke onafhankelijke RNAi constructen, zoals dsRNAs, weliswaar hetzelfde gen zullen uitschakelen, maar dat hun eventuele off-targets totaal verschillend zijn. Hoewel deze aanname gevoelsmatig correct lijkt te zijn, was dit eigenlijk nog nooit op een wetenschappelijke rekenkundige manier aangetoond. Hoofdstuk 3 brengt daar verandering in en de analyses daarin laten zien dat deze aanname lang niet altijd waar blijkt te zijn. Sterker nog, eigenlijk alle genen binnen het *Drosophila* genoom blijken veel sequenties te hebben die potentieel tot overlappende off-targets kunnen leiden.

Deze uitkomst was verrassend, maar is statistisch gezien heel logisch en goed te vergelijken met de zogenaamde verjaardags-paradox. Deze paradox beschrijft dat er slechts 23 willekeurige mensen in één ruimte nodig zijn om de kans waarop 2 mensen op dezelfde dag jarig zijn 50% te laten zijn, terwijl je dit aantal mensen intuïtief veel groter zou schatten. Het grote verschil tussen het gevoelsmatige aantal en het werkelijk aantal komt voort uit de manier waarop men de vraagstelling benadert. De vraag is namelijk niet: "Hoeveel mensen moeten er in een ruimte geplaatst worden, totdat er iemand is gevonden die dezelfde verjaardag heeft als ik?", maar de vraag is: “Hoeveel willekeurige mensen met willekeurige verjaardagen moeten er in een ruimte geplaatst worden zodat de kans 50% is dat er twee bijzitten die op dezelfde dag jarig zijn?”. Dus ieder willekeurig persoon in die ruimte mag met ieder willekeurig andere persoon een gemeenschappelijke verjaardag hebben en dit vergroot de kans vele malen. Hetzelfde is van toepassing op de vraag hoe veel verschillende sequenties er nodig zijn om 'toevallig' dezelfde off-target te vinden. We moeten namelijk niet alleen kijken of de 2 unieke dsRNAs die de onderzoeker wil gebruiken voor een RNAi experiment toevallig ergens exact dezelfde sequenties hebben (dit is in deze context niet relevant), maar we moeten ook kijken of de 2 unieke dsRNAs een gemeenschappelijke off-target hebben. Dit kan gebeuren wanneer een gemeenschappelijke off-target bestaat uit bijvoorbeeld ABCD, terwijl dsRNA 1 AB heeft en dsRNA2 CD. De 2 dsRNAs zijn weliswaar uniek, maar ze kunnen toch beide aangrijpen op hetzelfde off-target gen en

dat is nu juist wat de wetenschapper wil voorkómen als er gebruik gemaakt wordt van dsRNAs. Naast de uitgebreide analyses die vooral te maken hebben met bovenstaande vraagstelling, laat Hoofdstuk 3 ook een oplossing zien voor Drosophila onderzoekers met behulp van een on-line computerprogramma om veel gerichter dsRNA's te ontwerpen die een lage kans hebben op een gemeenschappelijke off-target. Dus door gebruik te maken van de beschreven methoden in dit hoofdstuk kunnen dsRNA's ontworpen worden met unieke off-target profielen die veel geschikter zijn om een gecontroleerd RNAi experiment uit te voeren.

Gecontroleerde RNAi: de praktijk

De effectiviteit van een gecontroleerd en beheerst RNAi experiment heeft zich bewezen in Hoofdstuk 4. In dit onderzoek werd niet de gehele vlieg, maar een cultuur van cellen afkomstig uit *Drosophila* embryo's gebruikt. Deze zogenaamde S2 cellen werden behandeld met RNAi constructen. Deze RNAi constructen hebben we ontworpen met het door ons gemaakte programma om zoveel mogelijk off-target effecten uit te sluiten. Met de ontworpen RNAi constructen werd een specifiek gen uitgeschakeld. Dat specifieke gen is erg interessant omdat bij patiënten, die aan een specifieke neurodegeneratieve ziekte (PKAN) lijden, ditzelfde gen ook is uitgeschakeld. Als we met RNAi het gen uitschakelen in cellen, kunnen we beter begrijpen waarom de patiënten ziek zijn. Eerst zal nu in iets meer detail de ziekte PKAN behandeld worden.

PKAN: een neurodegeneratieve ziekte

PKAN is een relatief zeldzame neuronale ziekte en wordt veroorzaakt door een mutatie in een gen. Dit gen wordt pantothenate kinase (PANK) genoemd en is essentieel voor een evolutionair zeer geconserveerde biochemische route. Deze route begint bij de verwerking van vitamine B5 en eindigt bij Coenzym A (CoA). Doordat deze biochemische route zo goed is geconserveerd binnen vrijwel alle organismen (van bacterie tot mens) is deze ziekte goed te bestuderen met behulp van de fruitvlieg. Met name door de eenvoud en snelheid waarmee de fruitvlieg gekweekt kan worden en de uitgebreide mogelijkheden om neuronale afwijkingen te induceren en te bestuderen is Drosophila een zeer geschikt modelorganisme. Naast studies in het hele organisme, zijn er ook Drosophila weefselkweek cellijnen waarin RNAi zeer effectief toegepast kan worden. We hebben eerst het PANK gen uitgeschakeld met behulp van RNAi, omdat het bestuderen van cellen

meestal eenvoudiger is dan het bestuderen van intacte organismen. Toen dit gelukt was, hebben we gekeken wat de consequenties hiervan zijn en vervolgens geprobeerd om de cellen met behulp van synthetische stofjes weer “beter” te maken. Een consequentie van het uitschakelen van het PANK gen is dat de cellen minder groeien en dat de CoA niveaus omlaag gaan. Nadat we het stofje pantethine toegevoegd hadden aan de cellen, bleek dat deze weer met de normale snelheid konden groeien en CoA niveaus werden hersteld. Ons onderzoek suggereert dat Pantethine ergens onderaan de CoA route kan instappen en dat op die manier onafhankelijk van het PANK gen, het CoA toch gemaakt kan worden. Ons onderzoek laat ook zien dat in de S2 cellijn pantethine beschermend werkt als het PANK gen defect is. Er bestaat een *Drosophila* model voor de ziekte PKAN en deze *Drosophila* mutante vliegen vertonen ernstige neurodegeneratie en hebben een zeer verkorte levensduur. Het pantethine werkt ook zeer beschermend tegen deze ziekteverschijnselen van het *Drosophila* PKAN model. In hoofdstuk 4 wordt dit onderzoek, dat gebaseerd is op een succesvol en gecontroleerd RNAi experiment en waarbij de tools worden gebruikt zoals ontworpen in hoofdstuk 2 en 3, uitvoerig beschreven. Er wordt ook bediscussieerd wat dit onderzoek zou kunnen betekenen voor patiënten die lijden aan PKAN.

miRNA’s en hun verwantschap aan siRNAs

De natuur kent nog een ander fenomeen dat nauw verwant is aan RNAi: de microRNA's (miRNA's). RNAi concentreert zich rond 21-bp sequenties terwijl miRNAs op nog veel kortere sequenties werken van rond de 6 basen. Het cel-eigen (endogene) miRNA proces speelt een belangrijke rol bij de regulatie van genen. Het is mogelijk om met één enkele miRNA grote groepen genen tegelijk te reguleren. Hoe dit proces precies wordt aangestuurd is nog niet geheel bekend, maar waarschijnlijk worden meerdere miRNA’s tegelijk ingezet om de genexpressies te dirigeren op een fijn niveau net zoals vele individuele muzieknoten een complex muziekstuk kunnen beschrijven. De vondst van miRNA's bracht meteen ook de eerdergenoemde term 'junk-DNA' tot wankelen: de miRNA's blijken namelijk vooral uit die stukken junk-DNA voort te komen en dus zijn deze stukken DNA geen 'junk' maar juist heel belangrijk. Hoewel het RNA afbreekproces door miRNA's en RNAi veel op elkaar lijken, kunnen ze onafhankelijk van elkaar voorkomen. Het ene proces kan namelijk chemisch of genetisch uitgeschakeld worden zonder dat het andere proces daar hinder

van ondervindt. Toch is er door de gemeenschappelijke component, namelijk het RNA, wel degelijk een overloop mogelijk tussen de beide processen. Helaas kunnen daardoor ook bij miRNAs off-targets voorkomen en dus moet ook daar rekening mee worden gehouden bij het bestuderen van de rol van miRNAs in de cel. Hoofdstuk 2 gaat daar verder op in en beschrijft middelen om naast de potentiële RNAi off-targets ook eventuele miRNA off-targets in ogenschouw te nemen. Het algoritme daarachter, het opzoeken van korte 6-bp sequenties, bleek ook zeer geschikt te zijn voor analyses in het humane genoom waar miRNA's in overvloed aanwezig zijn. Hoofdstuk 5 demonstreert dat aan de hand van een casus waarbij de betrokkenheid van miRNAs in Hodgkin's lymphoma (HL; witte bloed cel kanker) wordt geanalyseerd. Naast het feit dat gemuteerde genen de oorzaak kunnen zijn van kanker (de zogenaamde oncogenen), is het namelijk ook mogelijk dat niet het gen zelf gemuteerd is, maar dat de regulatie van een gen verstoord is. miRNA's kunnen daar de boosdoeners van zijn want er is aangetoond dat miRNAs ontregeld kunnen raken en daarmee ook de betrokken genen kunnen ontregelen. hoofdstuk 5 vergelijkt het miRNA profiel van gezonde cellen met het miRNA profiel van HL kankercellen. Er bleken inderdaad specifieke verschillen te zijn, maar alleen het aantonen van het verschil tussen de 2 cellijnen is niet voldoende bewijs om het kanker fenotype van de HL cellen te verklaren. Het door ons ontworpen RNAi-Select programma werd daarom ingezet en daaruit bleek dat de miRNA sequenties inderdaad terug te vinden zijn in de betrokken oncogenen in de HL cellijnen. Daardoor kon ook op sequentieniveau aangetoond worden dat die miRNAs waarschijnlijk betrokken zijn bij HL. Deze additionele bioinformatische analyse kon verder ondersteuning bieden bij de natte experimenten die het uiteindelijke onomstoten bewijs konden leveren dat de miRNA’s waren betrokken bij het (ont)regelen van bepaalde oncogenen. Bovendien werd een algemene methode ontwikkeld om ook de betrokkenheid van andere miRNAs in cellijnen aan te tonen op een manier waarop dat nog niet eerder mogelijk was.

Conclusie

Dit proefschrift gaat met name over RNA sequenties en de daaraan gerelateerde eiwitregulatiemechanismen. Door de relatief korte sequenties waarmee deze mechanismen werken kan de specificiteit in het geding komen. De natuur kan dat zelf uitstekend oplossen door gebruik te maken van de zogenaamde 'spatiele' (ruimtelijk) en 'temporele' (in de tijd)

verschillen. Met andere woorden, een organisme kan een miRNA op een heel specifiek tijdstip activeren tijdens de ontwikkeling en dan ook nog eens alleen in bepaalde weefsels. Indien miRNAs echter op verkeerde tijdstippen en/of in 'vreemd' weefsel tot expressie komen, dan kan dit onnatuurlijke off-targets veroorzaken. De RNAi technologie is ook een onnatuurlijke situatie: er wordt geen rekening gehouden met tijd en/of ruimte waarin de lichaamsvreemde moleculen ingebracht worden in het levende systeem. We hebben aangetoond dat er een onvermijdelijke overlap bestaat met andere sequenties in het genoom en dat kan consequenties hebben voor het succes van de experimenten die met RNAi technologie zijn uitgevoerd. Indien hier wel degelijk rekening mee wordt gehouden zoals wij aannemelijk hebben gemaakt met het RNAi-Select programma, is de techniek echter nog krachtiger dan ooit en blijkt het heel goed mogelijk te zijn om met potentiële off-targets om te gaan. We hebben voor de werking van RNAi-Select een database gemaakt met uitgebreide informatie over het genoom van *Drosophila* met in die database alle mogelijke off-targets in kaart gebracht. Ook hebben we, met behulp van de gebruikte algoritmes, RNAi technieken geoptimaliseerd en mogelijke miRNA targets beter kunnen aanwijzen. Dit heeft weer geleid tot het verder oplossen van medische gerelateerde vragen. Na dit werk begrijpen we beter hoe een specifieke vorm van leukemie ontstaat en hebben we een mogelijke basis voor het ontwikkelen van een therapie van een zenuwziekte waarvoor tot op heden nog geen enkel medicijn beschikbaar is.

Dankwoord

Geen enkele promotie wordt alleen gedaan en daarom is ook geen proefschrift compleet zonder een dankwoord. Mensen van het ‘lab’, vrienden en familie hebben allen een belangrijke rol gespeeld in mijn promotie en die wil ik op deze plek in het bijzonder bedanken.

Ody, voor jou was ik wellicht een beetje vreemde AIO eend in de bijt. Hoewel ik was begonnen met het vliegenwerk, verschoof mijn focus al snel naar de wereld van de bio-informatica. Je hebt mij de vrijheid geschonken om datgene te doen wat echt mijn passie was. Ik ben mij er terdege van bewust dat ik mijzelf gelukkig mag prijzen met een promotor die de zelfontwikkeling van haar AIO's een hoge prioriteit geeft. Ik heb de werkbesprekingen altijd zeer gewaardeerd en kijk ook met veel plezier terug op het 'avontuur' dat we hebben gehad met de diverse wetenschappelijke bladen. Naast je eigen visie en inbreng heb je een team van de beste mensen kunnen samenstellen waardoor ik mijn promotie met succes af kon ronden. Bedankt voor alles!

Ritsert, jij bent halverwege ingehaakt om de bio-informatica kant van mijn promotie te begeleiden. Hoewel je aan de top staat van de wetenschap en daardoor een overvolle agenda hebt en vaak elders op de wereld was te vinden, heb ik nooit op je hoeven wachten en heb je altijd de tijd genomen om met steengoede raad en daad mij (ons) bij te staan. Ook bij jou ben ik mij er terdege van bewust dat ik mij gelukkig mag prijzen met jou als promotor. Tezamen zijn we een lang (!) traject ingegaan, en hoewel het uiteindelijk niet die ene Cell of Nature publicatie is geworden is het werk uiteindelijk toch op een mooie plek terechtgekomen waar jij een belangrijke bijdrage aan hebt gehad.

Harrie, we hebben op een enkele uitzondering na nooit op wetenschappelijk gebied met elkaar gesproken, maar meestal over hele andere zaken. Het was een welkome afwisseling op de alledaagse werkzaamheden en heb altijd met heel veel genoegen die gesprekken met je gevoerd!

Mijn direct kamergenoten, ook jullie bedankt! De werksfeer was zeer prettig en de uitjes waren altijd gezellig. Ik zal daar altijd met veel plezier aan terugdenken! Floris, Anil, Kasia en Xia, you guys are the best roommates I could wish for. We had tons of fun with stuff that probably nobody

understands, and that just proves that we had a special collegial bond that probably most other people at the floor were jealous about. This created an atmosphere that allowed me to enjoy working every day. I could literally double the size of my thesis with nice anecdotes. Let's keep in touch over a coffee, Dalwhinnies, or even some weird Chinese tea shall we?

Er zijn tijdens de 5 jaar die ik heb doorgebracht bij SSCB veel mensen gekomen en ook weer gegaan. Willy, bedankt voor de korte samenwerking aan het begin van mijn promotie. Jurre, Maria, Marianne en Michel; bewoners van 'de andere kamer', bij jullie was het altijd erg gezellig om tijdens de koffiegang even langs te gaan en om tijdens de PCR's en blotwerk een praatje mee te maken. Jurre, wij hebben ook na je promotie nog altijd contact gehouden en ik hoop dat we die vriendschap nog lang zullen voortzetten!

Bedankt ook aan alle andere collega's van de afdeling SSCB. Bart, Hette, Jeanette, Maria, Martha, jullie zijn de zeldzame collega's die er van begin tot eind bij waren. Ik heb altijd met veel plezier de 11.30 lunch met jullie doorgebracht, en natuurlijk ook met de mensen die daarna nog bijschoven zoals Anne Jan, Marianne en anderen.

Ook wil ik nog de Stam Cell mensen aan het andere eind van de gang bedanken voor de leuke labdagen en de wandelgang gesprekken/discussies die her en der plaatsvonden.

Marike, bedankt voor je hulp om mijn Nederlandse samenvatting leesbaar en correct te maken!

Anke van den Berg en Lu Ping, het miRNA werk waar we aan hebben gewerkt was erg interessant en de samenwerking was zeer prettig, bedankt hiervoor.

Hans, bedankt voor alle hulp die je hebt geboden met betrekking tot de statistische berekeningen. Je bijdrage en inzicht is zeer gewaardeerd en was een belangrijk stap richting de acceptatie van het stuk!

Wouter, wij hebben reeds een lange geschiedenis met samenwerken. Onze samenwerking is ooit begonnen met het 'binnentikken' van posterpresentaties en 'Young Investigator Award'-achtige uitreikingen, en nu zijn we zelfs een gezamenlijk Bio-ICT bedrijf begonnen. Onze

vriendschap en toewijding om samen iets succesvols te maken is voor mij van grote waarde en het geeft mij een genoegen dat jij op de grote promotie dag als paranimf naast mij gaat staan.

Henk (vader) & Willy (mem), ook jullie hebben een belangrijke bijdrage geleverd in waar ik vandaag de dag sta. In het bijzonder dank ik jullie voor jullie toewijding en stimulering, die ervoor heeft gezorgd dat ik altijd verder ben gaan kijken dan mijn eigen neus lang is en nooit op mijn lauweren ben gaan rusten. Hoewel ik nog maar 1 paranimf mocht hebben, zie ik Henk als symbool voor jullie beiden om de belangrijke promotie dag naast mij door te brengen.

Petra, ook jij hebt mij altijd gestimuleerd en bent blijven geloven in al mijn ondernemingen die ik heb gedaan sinds wij elkaar kennen. Jouw input had soms een cruciale bedrage aan mijn werk en de discussies vormden een voedingsbodem voor nieuwe ideeën. Ik prijs mij ziels (souly) gelukkig met een vrouw zoals jij. Wij gaan nu samen onze eigen onderneming in met ons kind op komst, iets waar ik zeker van ben dat wij daar met volle tuigen van gaan genieten!

Curriculum Vitae

Personal Details

Name: Erwin Seinen
Place of Birth: Leeuwarden, Netherlands
Date of Birth: 28 February 1978

Education & Positions

1999-2005: Master study Biology, University of Groningen
Specialization: Molecular and medical genetics

2001-current: CEO (DGA) Winkelplein.nl

2005- 2010: PhD student (AIO), radiation and stress cell biology
University Medical Centre Groningen

2010-current: Product Manager/Co-owner of Bio-ITech

Honors and awards

Young Investigators Award (2002).

List of Publications

- Seinen E, Burgerhof JGM, Jansen RC, Sibon OCM (2010) RNAi Experiments in D. melanogaster: Solutions to the Overlooked Problem of Off-Targets Shared by Independent dsRNAs. PLoS ONE 5(10): e13119. doi:10.1371/journal.pone.0013119.
- L. P. Tan, **E. Seinen**, G. Duns, D. de Jong, O.C.M. Sibon, K. Kok, B.J. Kroesen, S. Poppema, A. van den Berg (2009). A high throughput experimental approach to identify miRNA target genes in Hodgkin lymphoma. *Nucleic Acids Res* 37: e373.
- A. Rana, **E. Seinen**, K. Siudeja, R. Muntendam, B. Srinivasan, J van der Want, S. Hayflick, D. Reijngoud, O. Kayser, O.C.M. Sibon (2010). Pantethine rescues a *Drosophila* model for Pantothenate Kinase-Associated Neurodegeneration. *Proc Natl Acad Sci U S A* 107: 6988-6993